MULTIMODAL ANALYSIS:
Informed Content Estimation and Audio Source Separation

Gabriel MESEGUER BROCAL
09/01/2017 – 09/07/2020

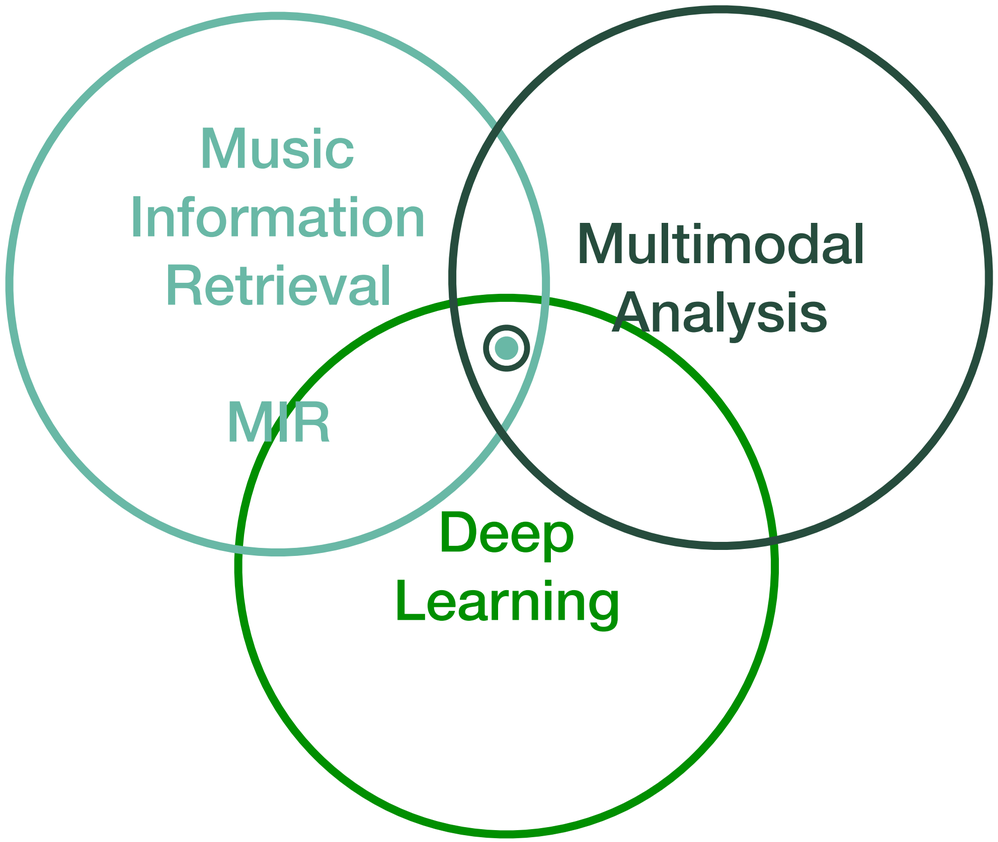
Directed by Geoffroy PEETERS

Introduction

It aims to understand and make of **music** data.

by ***researching***, ***developing*** **computational** systems to solve music related problem.

It is a **multidisciplinary** field with theories, concepts and techniques from music, computer science, signal processing and cognition.



Meinard Müller (2015), **Fundamentals of Music Processing**

Schedl, M., Gómez Gutiérrez, E., & Urbano, J. (2014). **Music information retrieval: Recent developments and applications.** Foundations and Trends in Information Retrieval.

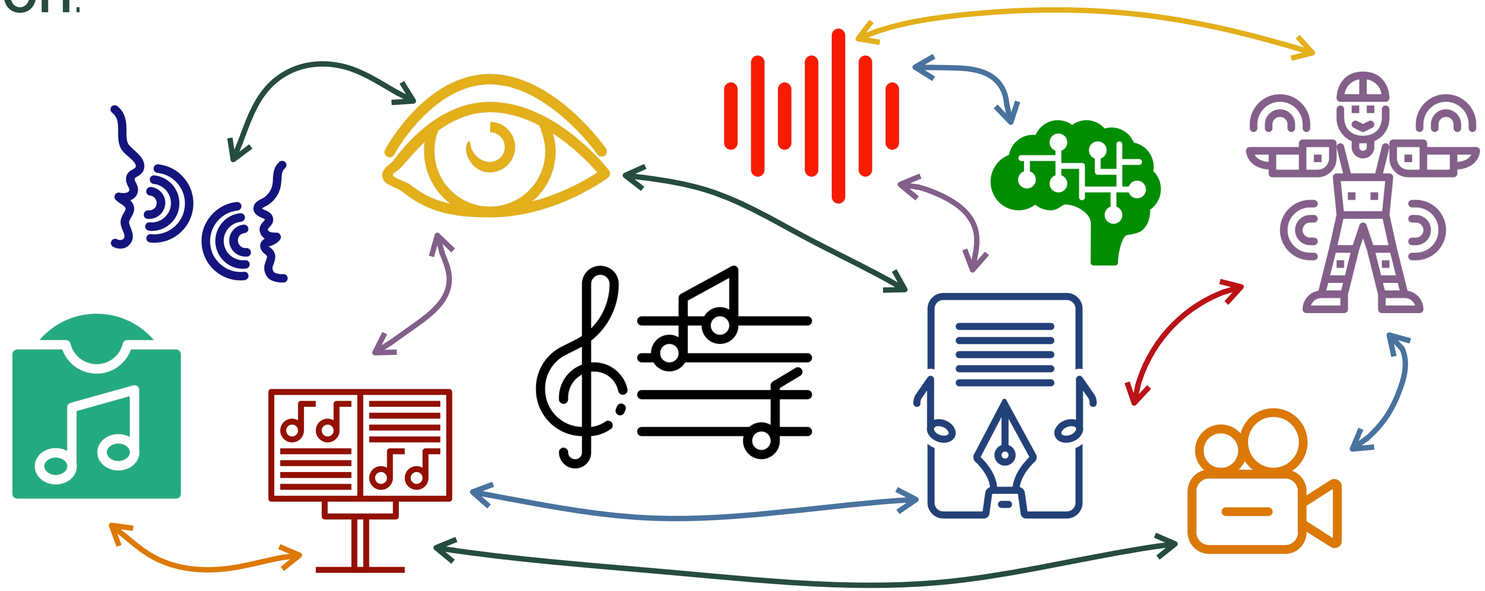
Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011, January). **Multimodal deep learning**. In *ICML*.

What is multimodal analysis?

In reality, **phenomena** are expressed through many different **domains**.

Our **understanding** of involves the **fusion** of different modalities into a **joint representation**.

Music is intrinsically multimodal



Icons from Noun Project

Multimodal analysis is the **machine learning** discipline that studies the interaction between modalities.

Goal: to improve model's performance.

Essid, Slim and Richard, Gaël (2012). ***Fusion of multimodal information in music content analysis***.

What is multimodal analysis?

In reality, **phenomena** are expressed through many different **domains**.

Our **understanding** of involves the **fusion** of different modalities into a **joint representation**.

These interaction can be described following **three** questions:

- **What**? - data/domain.
- **How**? - methodology to develop models.
- **When/where**? - data and methodology.

Multimodal analysis is the **machine learning** discipline that studies the interaction between modalities.

Goal: to improve model's performance.

Bengio, Yoshua and Courville, Aaron and Vincent, Pascal (2013). ***Representation Learning: A Review and New Perspectives***. IEEE Transactions on Pattern Analysis and Machine Intelligence

Our formalisation

Motivation: enlarge the music analysis by adding other domains to the audio signal and to improve **MIR** tasks.

Singing voice

Why? Central element in popular music:

- As musical instrument: lead melody / structures.
- Add semantic meaning: tells stories / conveys emotions.

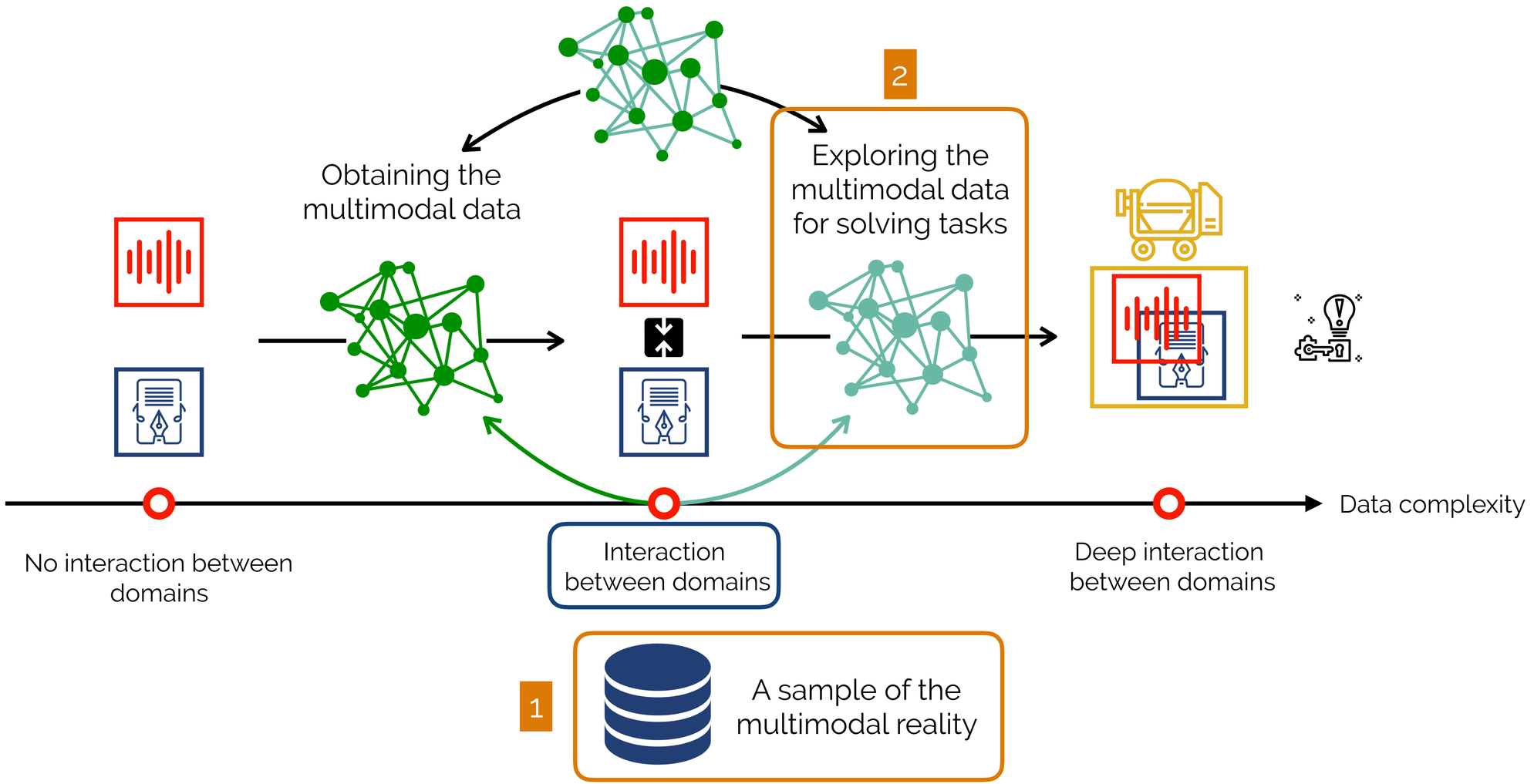
What: audio + lyrics

When: aligned in time → time synchronisation = direct interaction.

How: deep neural networks → great results, flexible, constant evolution, tools, …

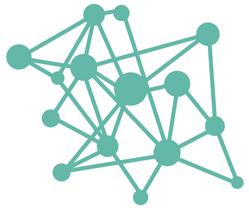
Demetriou, Andrew et all. (2018). ***Vocals in Music Matter: the Relevance of Vocals in the Minds of Listeners.*** In ISMIR 2018, Paris, France.

Overview


2
Obtaining the multimodal data
Exploring the multimodal data for solving tasks
Data complexity
No interaction between domains
Interaction between domains
Deep interaction between domains
1
A sample of the multimodal reality

Plan

1. Introduction
2. Dataset of Aligned Lyric Information - DALI
   - 2.1. Motivation
   - 2.2. Creation
   - 2.3. Training with noisy data
3. Multimodal tasks
   - 3.1. Structures analysis
   - 3.2. Source separations
     - 3.2.1.1. Multitasks
     - 3.2.1.2. Vocals
4. Conclusions and future work




DALI

2.1

Motivation

What kind of **data** do we need for carry on our research?

Is there any dataset that much our needs?

2.1

Motivation

Is there any dataset that much our needs? NO

Dataset	Number of songs	Language	Audio type	Granularity
(Iskandar et al., 2006)	No training. 3 tests songs	English	Polyphonic	Syllables
(Wong et al., 2007)	14 songs divided into 70 segments with 20s long	Cantonese	Polyphonic	Words
(Müller et al., 2007)	100 songs	English	Polyphonic	Words
(Kan et al., 2008)	20 songs	English	Polyphonic	Section Lines
(Mesaros and Virtanen, 2010)	Training: 49 fragments ∼25 seconds for adapting a phonetic model Testing: 17 songs	English	Training: A Capella Testing: Vocals after source separation	Lines
(Hansen, 2012)	9 pop music songs	English	Both, Polyphonic A Capella	Words Lines
(Mauch et al., 2012)	20 pop music songs	English	Polyphonic	Words
DAMP dataset, (Smith, 2015)	34k **amateur** versions of 301 songs	English	Amateurs A Capella	Not time-aligned lyrics only textual lyrics
DAMPB dataset, (Kruspe, 2016)	A **DAMP** subset with 20 performances of 301 songs	English	Amateurs A Capella	Words Phonemes
(Dzhambazov, 2017)	70 fragments of 20 seconds	Chinese Turkish	Polyphonic	Phonemes
(Lee and Scott, 2017)	20 pop music songs	English	Polyphonic	Words
(Gupta et al., 2018)	A **DAMP** subset with 35662 segments of 10s long	English	Amateurs A Capella	Lines
$\text{Jamendo}_{\text{aligned}}$, (Ramona et al., 2008) (Stoller et al., 2019)	20 Creative commons songs	English	Polyphonic	Words

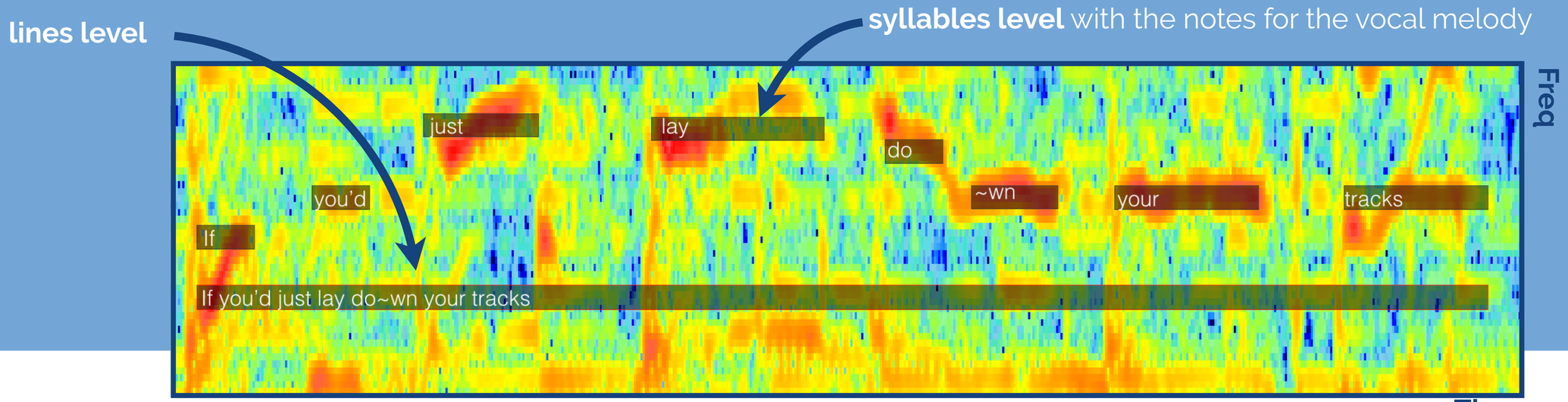
2.1

Motivation

Dataset of Aligned Lyric Information - ***DALI*** https://github.com/gabolsgabs/DALI

7756 songs, each one has:

- its **audio** in full-duration.
- its **time-aligned notes** for vocal melody.
- its **time-aligned lyrics** in four levels of granularity: ***syllables, words, lines and paragraphs***.



close-up of the spectrogram

2.2

Creation

Creating such dataset is a **complex** problem: time-consuming and expert knowledge.

Our solution: to adapt existing resources outside our community:

Karaoke
Open source communities

13339
annotations files

Issues:

1. The data is not standardised and there are some missing information.

2. Unknown audio → inverse scenario than usual annotation but no audio Vs audio and not annotations.

Creation

Solution to no standardisation and some missing information

Info: Artist name + Song title

```
#ARTIST:Dream Theater
#TITLE:Panic Attack
#MP3:Dream Theater - Panic Attack.mp3
#BPM:253
#GAP:50080
: 0 6 3 All
: 6 4 3 wound
: 12 4 2 up
: 20 5 2 on
: 26 3 2 the
: 29 3 0 ~
: 33 4 0 edge
- 38
: 40 6 3 Ter
: 46 4 3 ri
: 52 8 2 fied
- 62
: 77 4 3 Sleep
: 83 2 3 dis
: 89 4 2 turbed
: 97 2 2 rest
: 103 2 2 less
: 109 6 0 mind
- 116
: 118 2 3 Spet
: 124 4 3 ri
: 130 8 2 fied
- 140
```

Pseudo bpm

Offset

Time onset

Time duration

Text

Musical note
Ref note 0 = C3

Lyrics at different granularities levels

Note frequency

Hierarchical connection

```
"annotations": {
  "notes" [
      {...},
      {
        'text': 'wound',
        'freq': [311.1269, 311.1269],
        'time': (50.4917, 50.7290),
        'index': 1
      },
      {...},
      {
        'text': 'ter',
        'freq': [311.1269, 311.1269],
        'index': 6,
        'time': (52.5075, 52.8633)
      },
      {
        'text': 'ri',
        'freq': [311.1269, 311.1269],
        'index': 6,
        'time': (52.8633, 53.1004)
      },
      {
        'text': 'fied',
        'freq': [293.6648, 293.6648],
        'index': 6,
        'time': (53.2190, 53.6933)
      },
      {...}
  ],
```

```
"words": [
    {...}
    {
      'text': 'terrified',
      'freq': [293.6648, 311.1270],
      'index': 1,
      'time': (52.5075, 53.6933)
    },
    {...}
],
```

Word frequency range

```
"lines":[
    {
      'text': 'all wound up on the edge',
      'freq': [261.6255, 311.1269],
      'time': (50.1360, 52.3297),
      'index': 0
    },
    {
      'text': 'terrified',
      'freq': [293.6648, 311.1270],
      'index': 0,
      'time': (52.5075, 53.6933)
    },
    {...}
],
```

Text segment

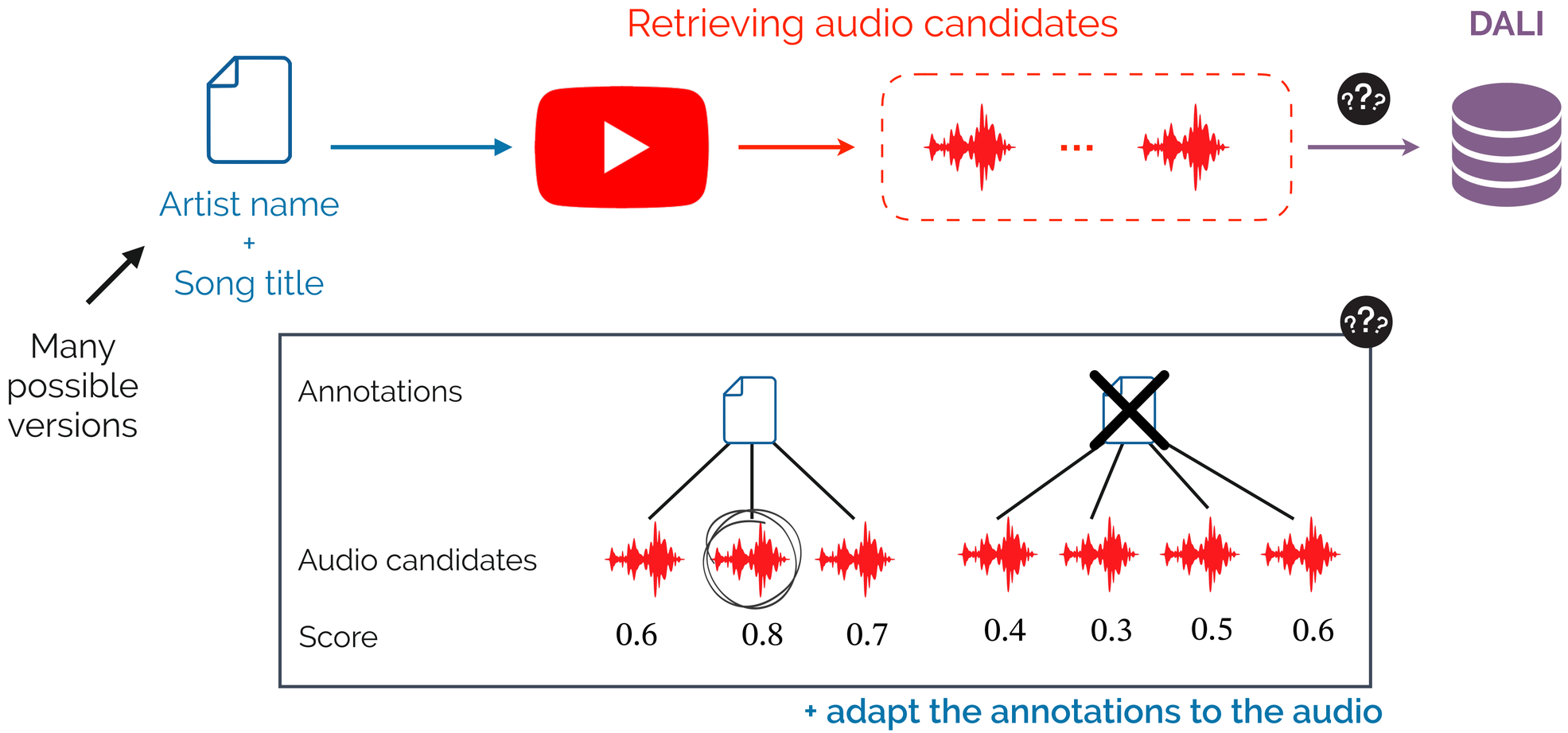
Text segment duration

```
"paragraphs":[
    {
      'text': 'all wound up on the edge terrified',
      'freq': [261.6255, 311.1270],
      'time': (50.1360, 53.6933)
    },
    {...}
]
```

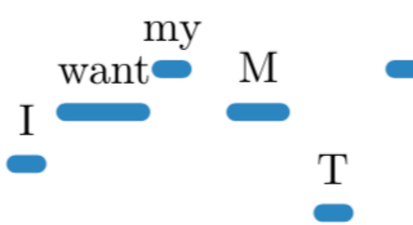
```
'phonemes': [
  {...},
  {
    'text': ['T', 'EH', 'R', 'AH', 'F', 'AY', 'D'],
    'freq': (293.6648, 311.1270),
    'index': 1,
    'time': (52.5075, 53.6933)},
  {...},
]
```

Creation

Solution to not knowing the audio:

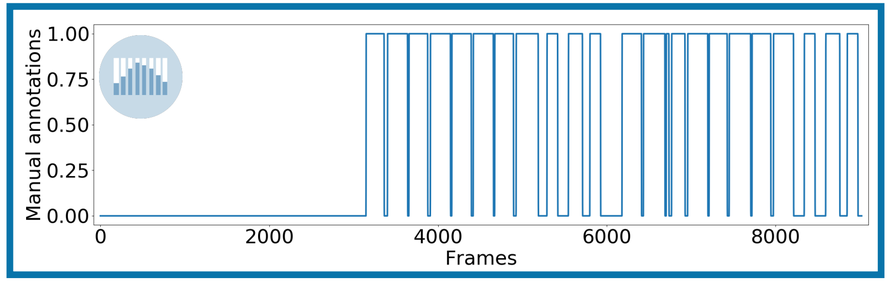


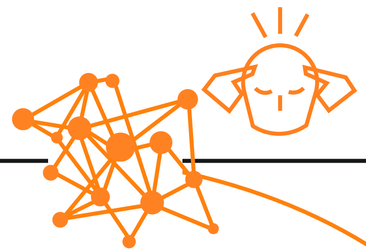
Creation



Voice Annotation Sequence

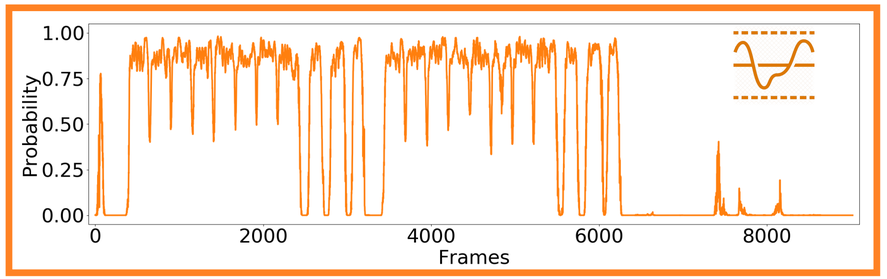
$$vas(t)$$



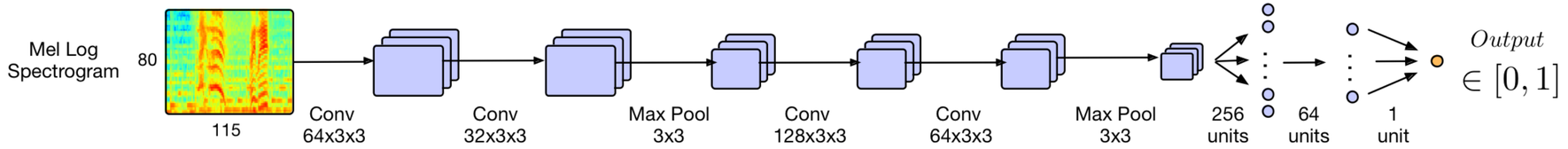


Singing Voice probability vector

$$\hat{p}(t)$$



Singing Voice detection system (SVD)

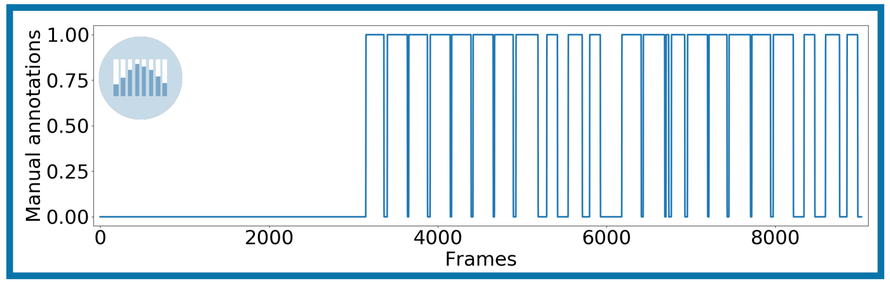


Schlüter, J. and Grill, T. (2015). ***Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks***. In ISMIR 2015, Malaga, Spain.

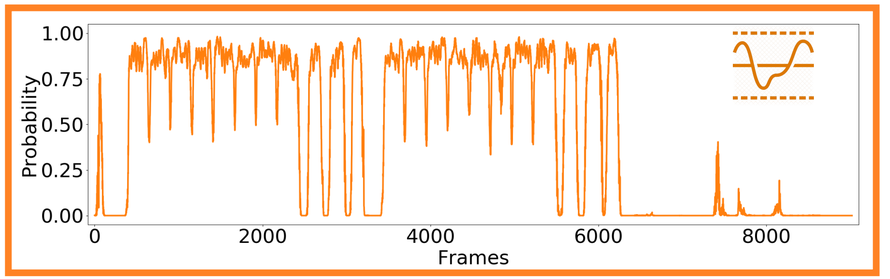
2.2

Creation

$vas(t)$



$\hat{p}(t)$



Hypothesis: if the annotations and the system that detects the singing voice are perfect, both vectors must be identical.

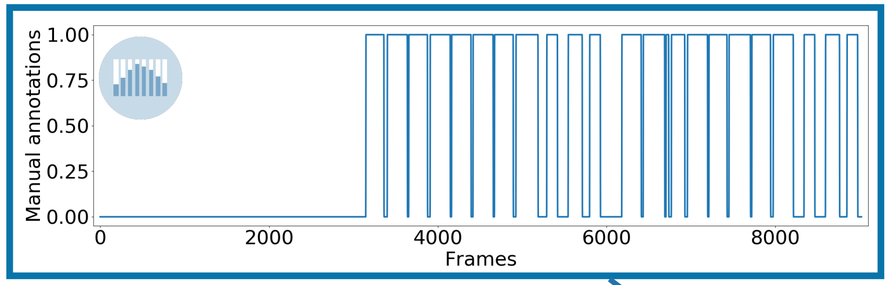
How do we compare them? Requirements:

- Fast: each entry has an average of 6 candidates.
- Normalised: being able to select candidates and filter possible errors.
- Do not modify the annotations just place them in the right position- for now we only want to find the audio, solving annotation errors is complex.

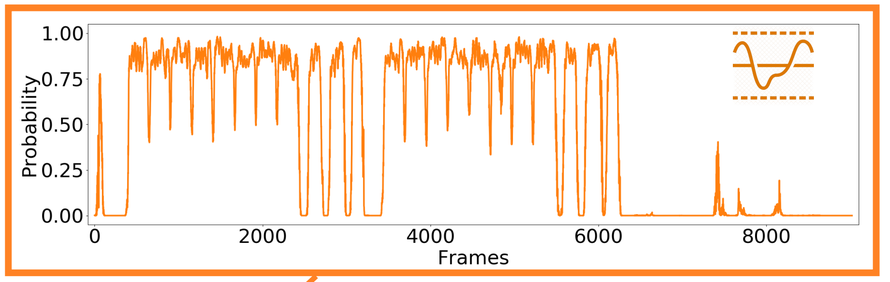
II

Creation

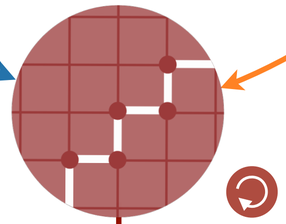
$vas(t)$

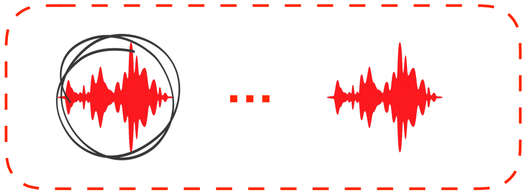


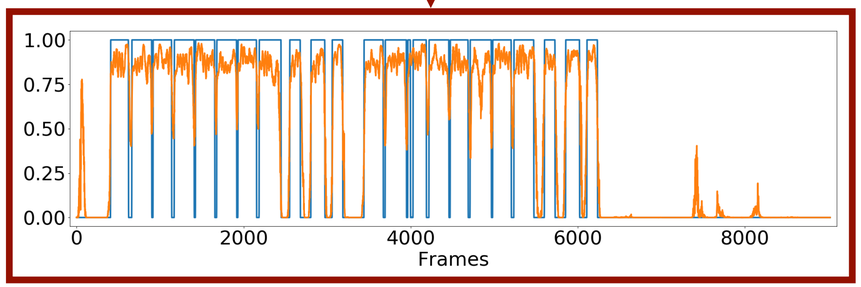
$\hat{p}(t)$



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







Alignment

$\hat{o}, \hat{fr}$

+

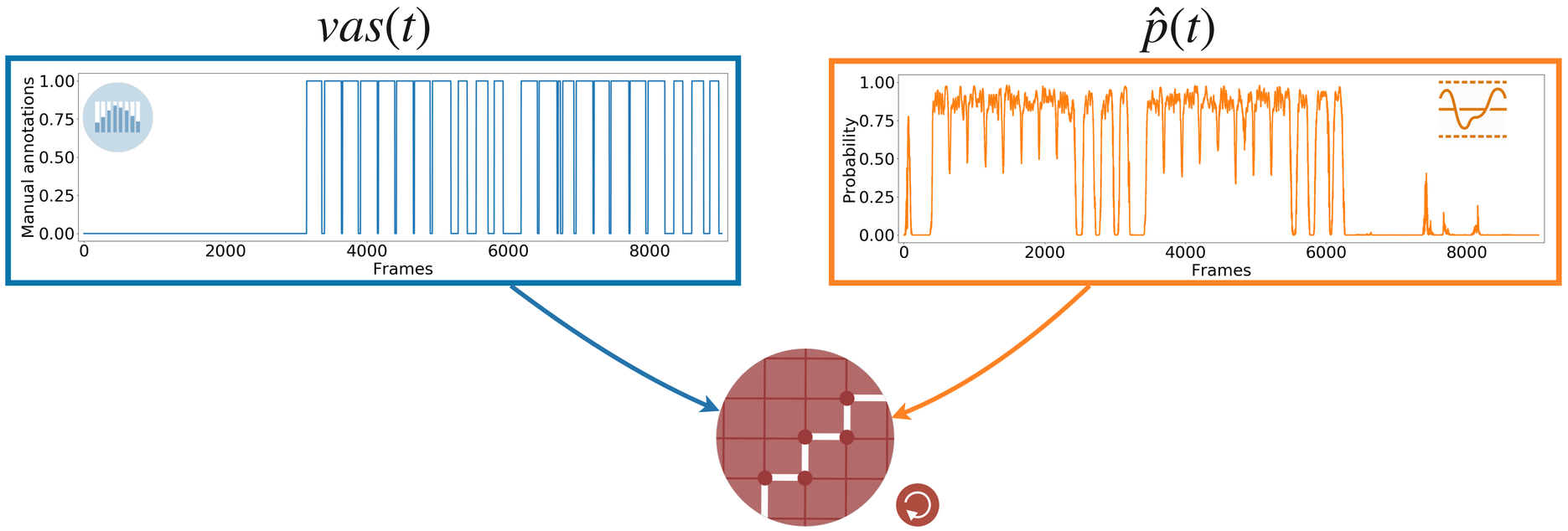
SCORE = 0.92

$NCC(\hat{o}, \hat{fr})$

Meseguer-Brocal, G. et all. (2018) **DALI: a large Dataset of synchronized Audio, LyrIcs and notes, automatically created using teacher-student machine learning paradigm**. (ISMIR).

II

Creation



This process strongly depends on the quality of . Small differences in similar score but very different alignments. **Need to improve the !!!**

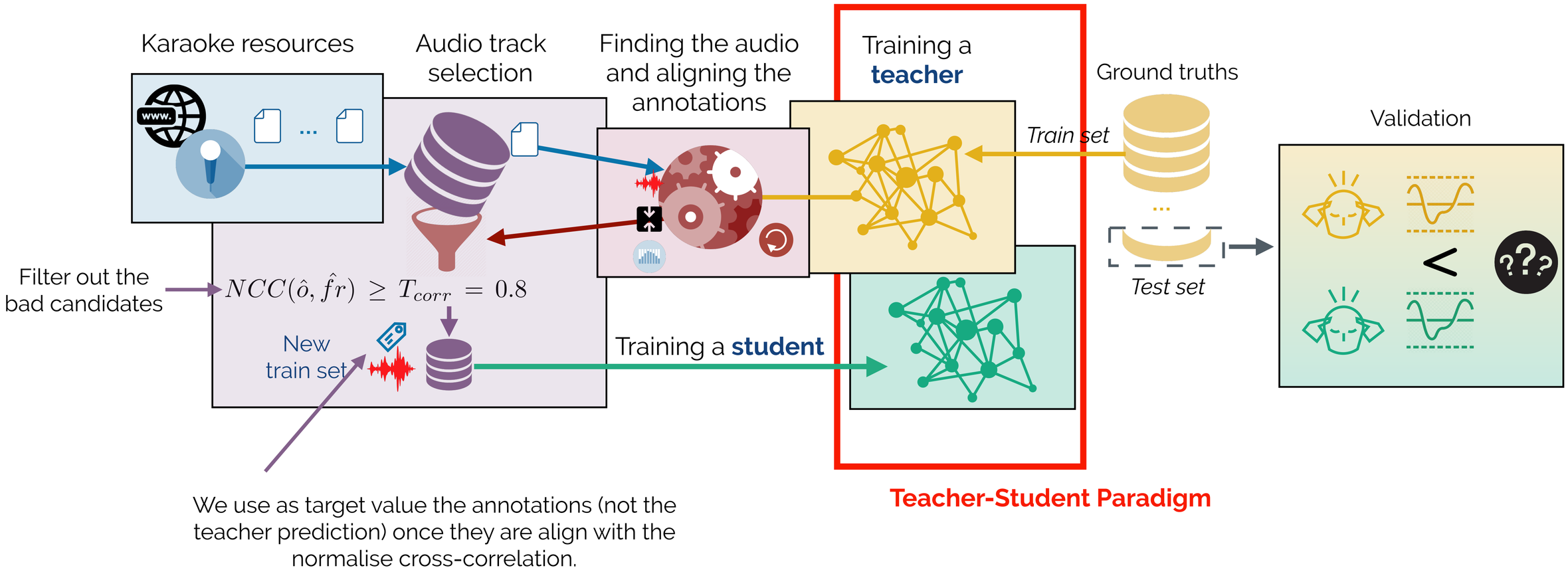
a. New architecture.

b. Train with better data.

Meseguer-Brocal, G. et all. (2018) **DALI: a large Dataset of synchronized Audio, LyrIcs and notes, automatically created using teacher-student machine learning paradigm**. (ISMIR).

2.2

Creation

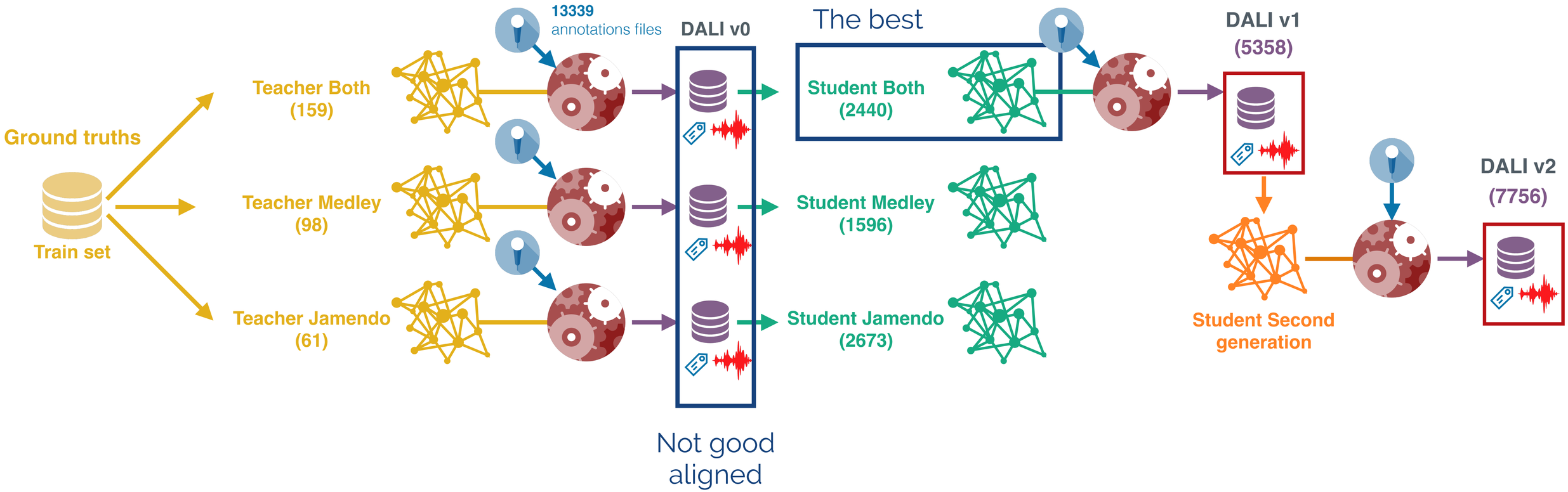


Gabriel Meseguer-Brocal, Alice Cohen-Hadria and Geoffroy Peeters. (2018) **DALI: a large Dataset of synchronized Audio, Lyrlcs and notes, automatically created using teacher-student machine learning paradigm**. (ISMIR).

Gabriel Meseguer-Brocal, Alice Cohen-Hadria and Geoffroy Peeters. (2020) **Creating DALI, a large dataset of synchronized audio, lyrics, and notes**. (TISMIR).

2.2

Creation



Gabriel Meseguer-Brocal, Alice Cohen-Hadria and Geoffroy Peeters. (2018) **DALI: a large Dataset of synchronized Audio, Lyrlcs and notes, automatically created using teacher-student machine learning paradigm.** (ISMIR).

Gabriel Meseguer-Brocal, Alice Cohen-Hadria and Geoffroy Peeters. (2020) **Creating DALI, a large dataset of synchronized audio, lyrics, and notes.** (TISMIR).

2.2

Creation

How to prove that the new vocal activity is better than the previous one?

Two experiments:

- **Singing voice detection**: binary frame accuracy after thresholding using the validation test set.

- **On alignment**: verify the accuracy of the  after finding the right alignment using 

For both experiments the each new **student** works better than its **teacher.**

Gabriel Meseguer-Brocal, Alice Cohen-Hadria and Geoffroy Peeters. (2018) **DALI: a large Dataset of synchronized Audio, LyrIcs and notes, automatically created using teacher-student machine learning paradigm**. (ISMIR).
Gabriel Meseguer-Brocal, Alice Cohen-Hadria and Geoffroy Peeters. (2020) **Creating DALI, a large dataset of synchronized audio, lyrics, and notes**. (TISMIR).

2.2

Creation

Singing voice detection

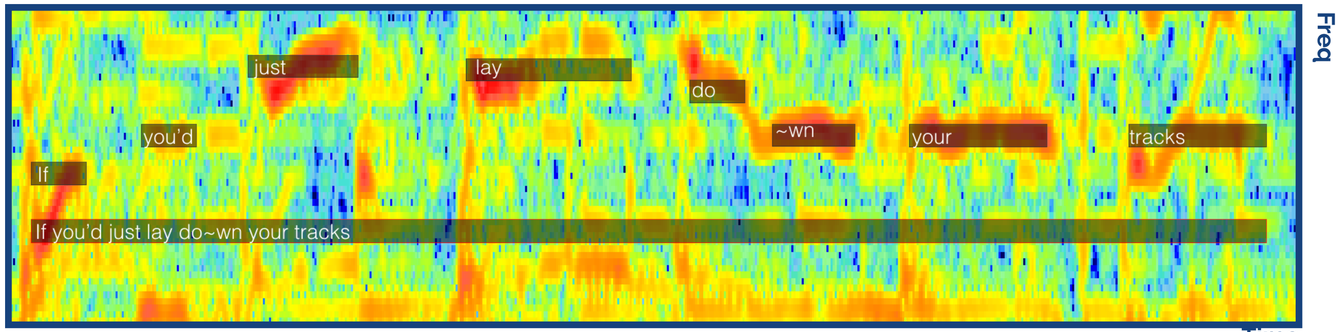
SVD system \ Test_set	J_test(16)	M_test(36)	**J_(test+train)(77)**	**M_(test+train)(98)**
T_J_train(61)	88.95% ± 5.71	83.27% ± 16.6	-	81.83% ± 16.8
S [T_J_train](2673)	87.08% ± 6.75	82.05% ± 15.3	87.87% ± 6.34	84.00% ± 13.9
T_M_train(98)	76.61% ± 12.5	84.14% ± 17.4	76.32% ± 11.2	-
S [T_M_train](1596)	82.73% ± 10.6	79.89% ± 17.8	84.12% ± 9.00	82.03% ± 16.4
T_both_train(159)	83.63% ± 7.13	83.24% ± 13.9	-	-
S [T_*J*+*M*_train](2440)	87.79% ± 8.82	85.87% ± 13.6	89.09% ± 6.21	86.78% ± 12.3
2G [T_S_*J*+*M*](5253)	**93.37% ± 3.61**	**88.64% ± 13.0**	**92.70% ± 3.85**	**88.90% ± 11.7**

Meseguer-Brocal, G. et all. (2018) **DALI: a large Dataset of synchronized Audio, Lyrlcs and notes, automatically created using teacher-student machine learning paradigm**. (ISMIR).
Meseguer-Brocal, G. et all. (2020) **Creating DALI, a large dataset of synchronized audio, lyrics, and notes**. (TISMIR).

Recap

V	Songs	Artists	Genres	Languages	Decades
1.0	5358	2274	61	30	10
2.0	7756	2866	63	32	10
Multitracks	512	247	32	1	7

Mixture = Vocals + accompaniment



close-up of the spectrogram

A sample of the multimodal reality

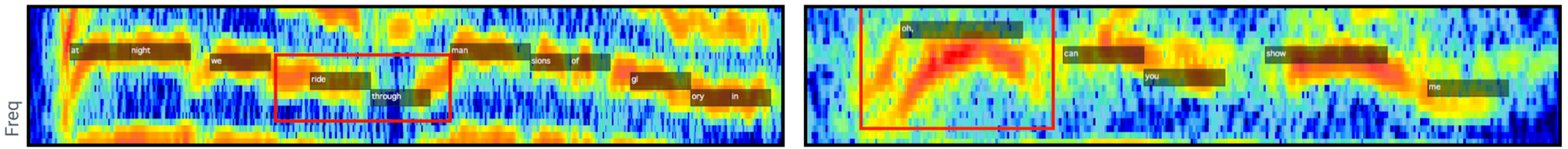
Gabriel Meseguer-Brocal, Alice Cohen-Hadria and Geoffroy Peeters. (2018) **DALI: a large Dataset of synchronized Audio, LyrIcs and notes, automatically created using teacher-student machine learning paradigm**. (ISMIR).

Gabriel Meseguer-Brocal, Alice Cohen-Hadria and Geoffroy Peeters. (2020) **Creating DALI, a large dataset of synchronized audio, lyrics, and notes**. (TISMIR).

2.3

Training with noisy data

Now we have: audio+annotations+metric of how good the annotations are but still with **errors**.



close-up of the spectrogram

2.3

Training with noisy data

Now we have: audio+annotations+metric of how good the annotations are but still with **errors**.

i. Can we automatically solve the errors?

2.3

Training with noisy data

Now we have: audio+annotations+metric of how good the annotations are but still with **errors**.

i. Can we automatically solve the errors? → Alignment techniques.

- Many different techniques and configurations.
- Quite sensible to changes in the process.

- Difficulty to quantify improvements or the lack of them.
- We **still** will not know where there are errors.

2.3

Training with noisy data

Now we have: audio+annotations+metric of how good the annotations are but still with **errors**.

i. Can we automatically solve the errors?
ii. Can we automatically find where there are errors?

2.3

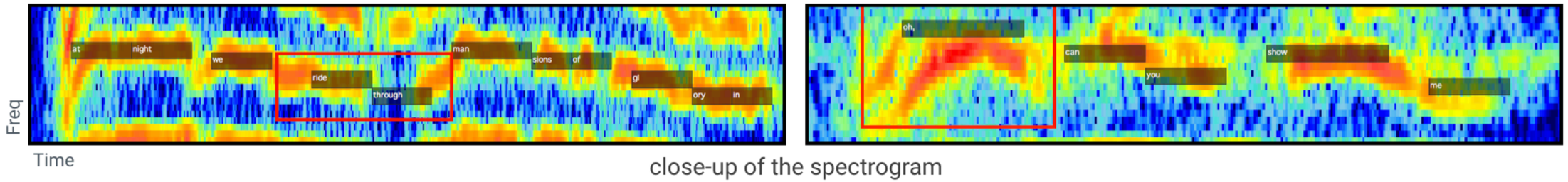
Training with noisy data

Now we have: audio+annotations+metric of how good the annotations are but still with **errors**.

i. Can we automatically solve the errors?

ii. Can we automatically find where there are errors?

Our hypothesis: **finding** errors is **simpler** than **solving** them for complex classification tasks with high numbers of classes.



close-up of the spectrogram

We test this idea for the **note** annotations only → errors in time and frequency.

2.3

Training with noisy data

Now we have: audio+annotations+metric of how good the annotations are but still with **errors**.

i. Can we automatically solve the errors?

ii. Can we automatically find where there are errors?

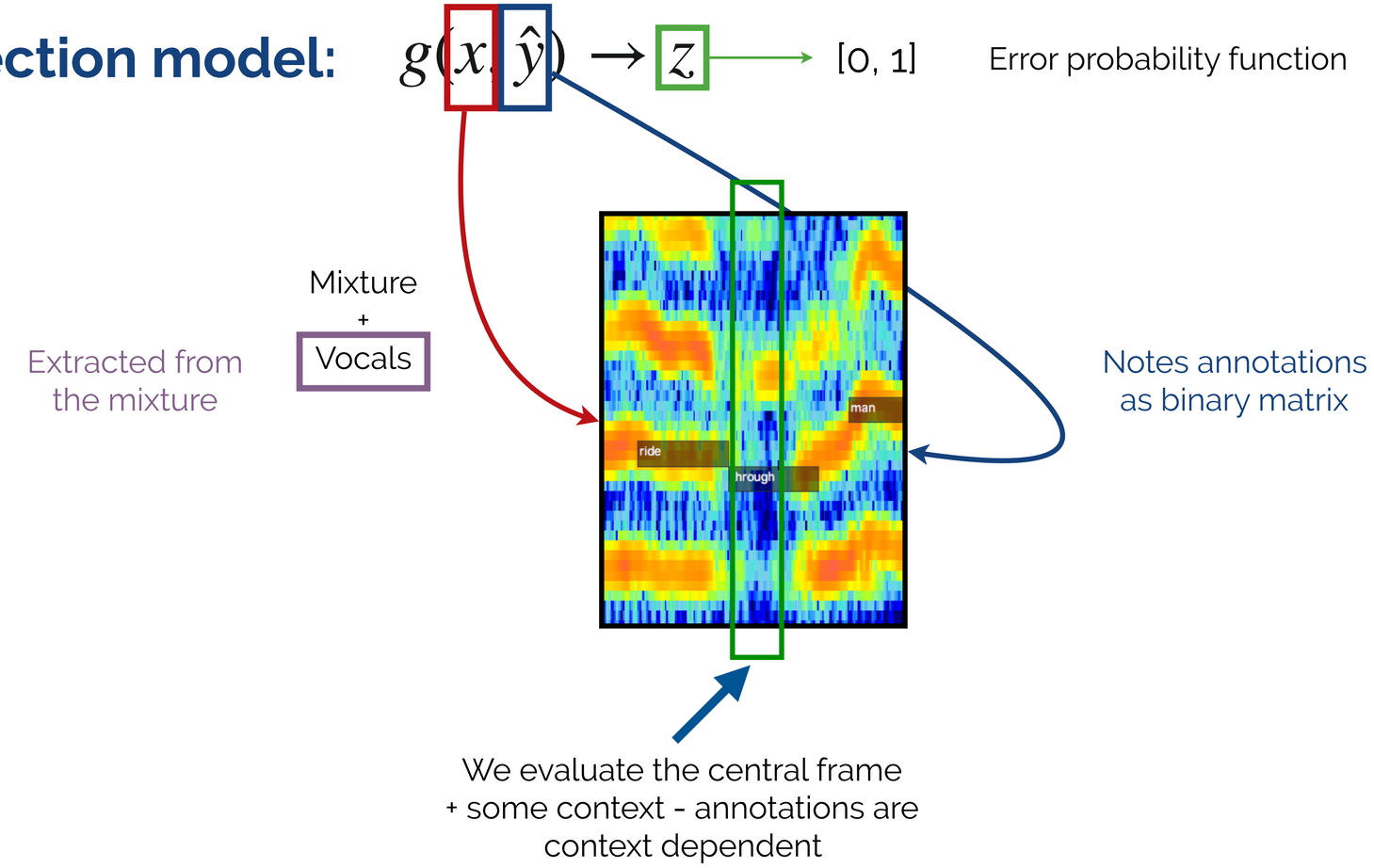
Data cleansing helps in detecting where the errors occur for:

1. Knowing the current status of the dataset.
2. Filtering for training.
3. Evaluate the alignment techniques to solve the errors in the annotations.

2.3

Training with noisy data

Error detection model: $g(x, \hat{y}) \rightarrow z$ → [0, 1] Error probability function

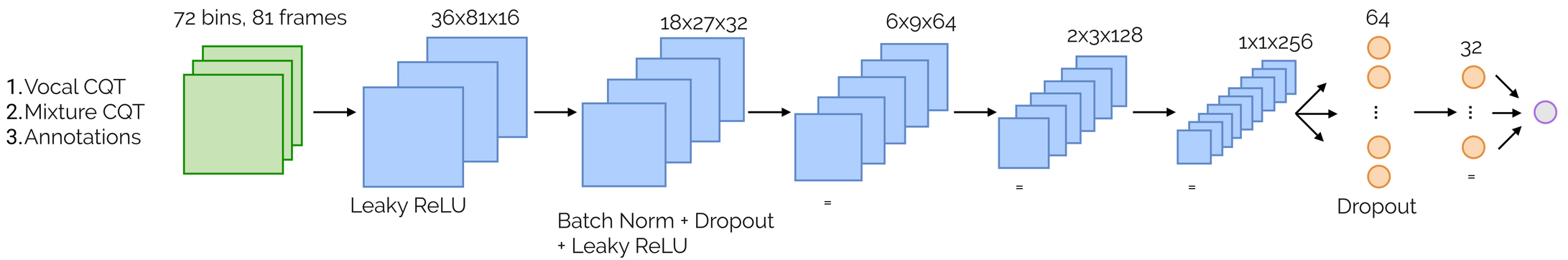


Gabriel Meseguer-Brocal, Rachel Bittner, Simon Duran, Brian Brost. (2020) **Self-Supervised Data Cleansing for Vocal Note Event Annotations.** (ISMIR) Under review

2.3

Training with noisy data

Error detection model: $g(x, \hat{y}) \rightarrow z$



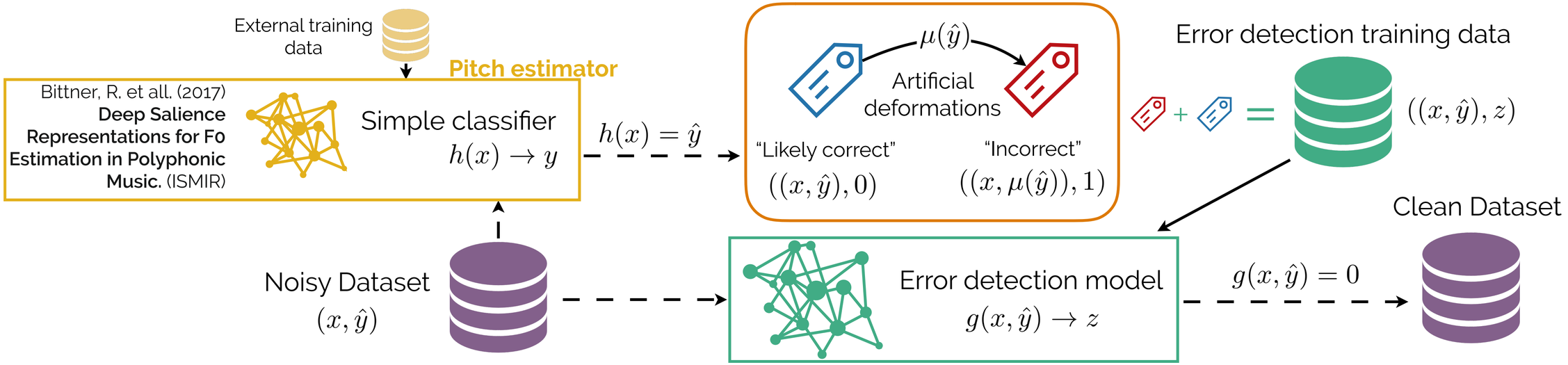
How to train this model if we do not know what annotations are good and what annotations are bad? z?

Gabriel Meseguer-Brocal, Rachel Bittner, Simon Duran, Brian Brost. (2020) **Self-Supervised Data Cleansing for Vocal Note Event Annotations.** (ISMIR) Under review

2.3

Training with noisy data

How to train this model if we do not know what annotations are good and what annotations are bad? z?

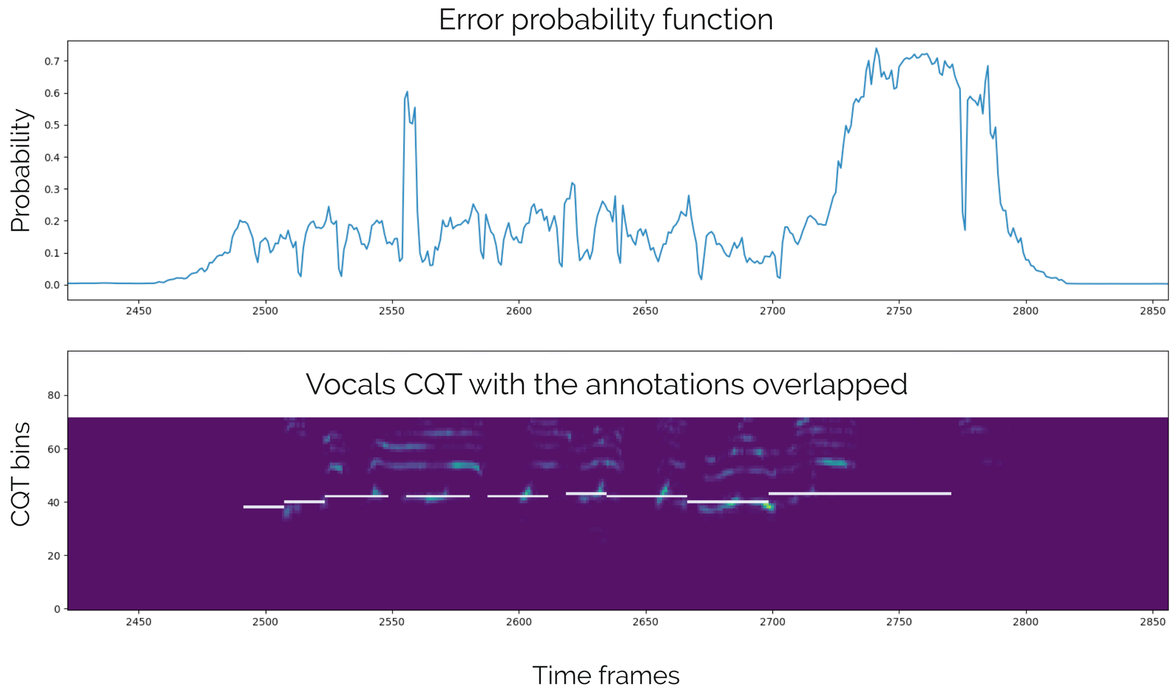


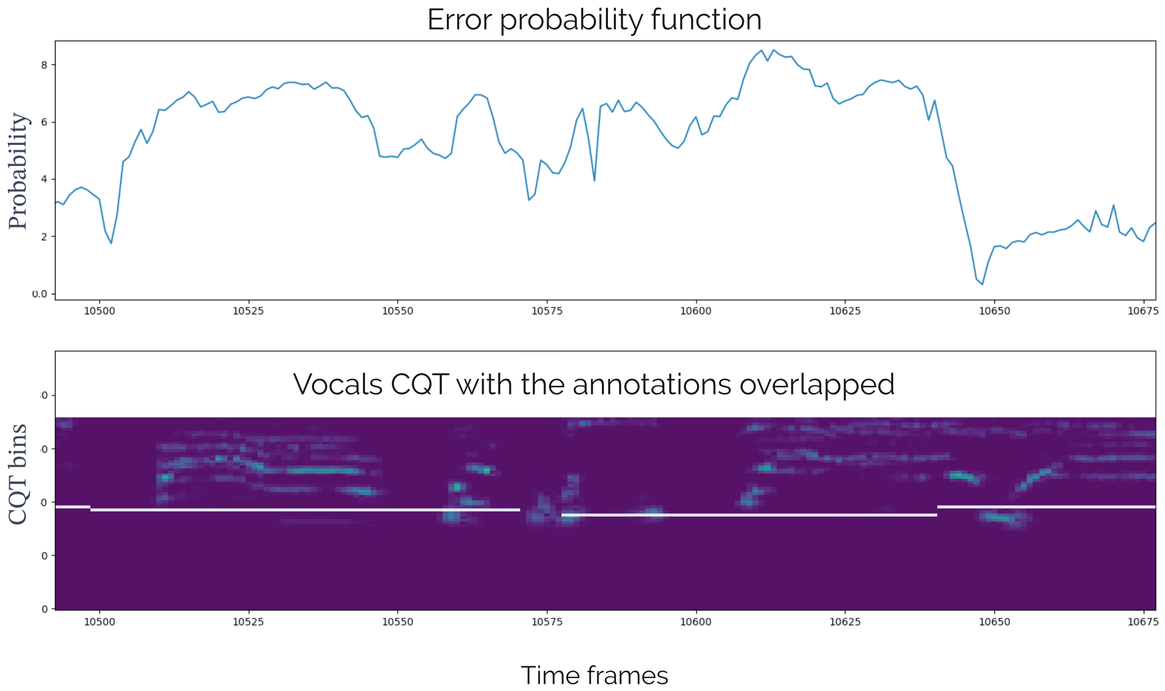
Gabriel Meseguer-Brocal, Rachel Bittner, Simon Duran, Brian Brost. (2020) **Self-Supervised Data Cleansing for Vocal Note Event Annotations.** (ISMIR) Under review

2.3

Training with noisy data

Error detection model: $g(x, \hat{y}) \rightarrow z$





Gabriel Meseguer-Brocal, Rachel Bittner, Simon Duran, Brian Brost. (2020) **Self-Supervised Data Cleansing for Vocal Note Event Annotations.** (ISMIR) Under review

2.3

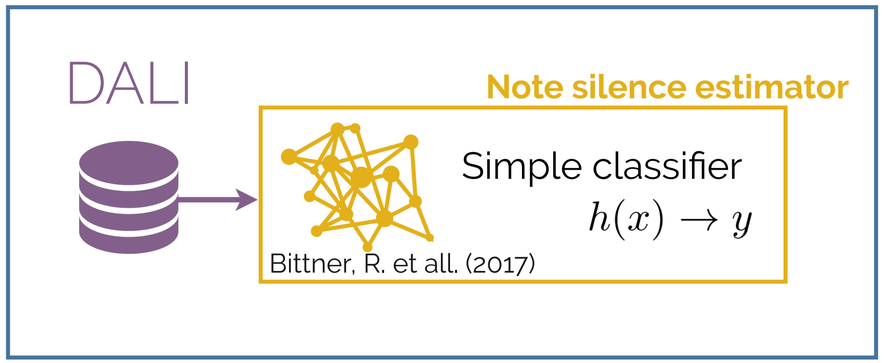
Training with noisy data

How to validate $g(x, \hat{y}) \rightarrow z$?

a. Directly: we do not have any "real" ground truth good and bad annotations (only likely correct and artificially created wrong examples).

b. Manually: it is infeasible (costly and required expert knowledge) and defeats the purpose of automating the process of correcting errors.

c. As in traditional data cleansing approaches → identifying incorrect annotations and remove them during training.



x3 - different versions of Dali

1. All
2. Filtered $z \leq 0.5$
3. Weighted loss by $1 - z$

Gabriel Meseguer-Brocal, Rachel Bittner, Simon Duran, Brian Brost. (2020)
Self-Supervised Data Cleansing for Vocal Note Event Annotations. (ISMIR)
Under review

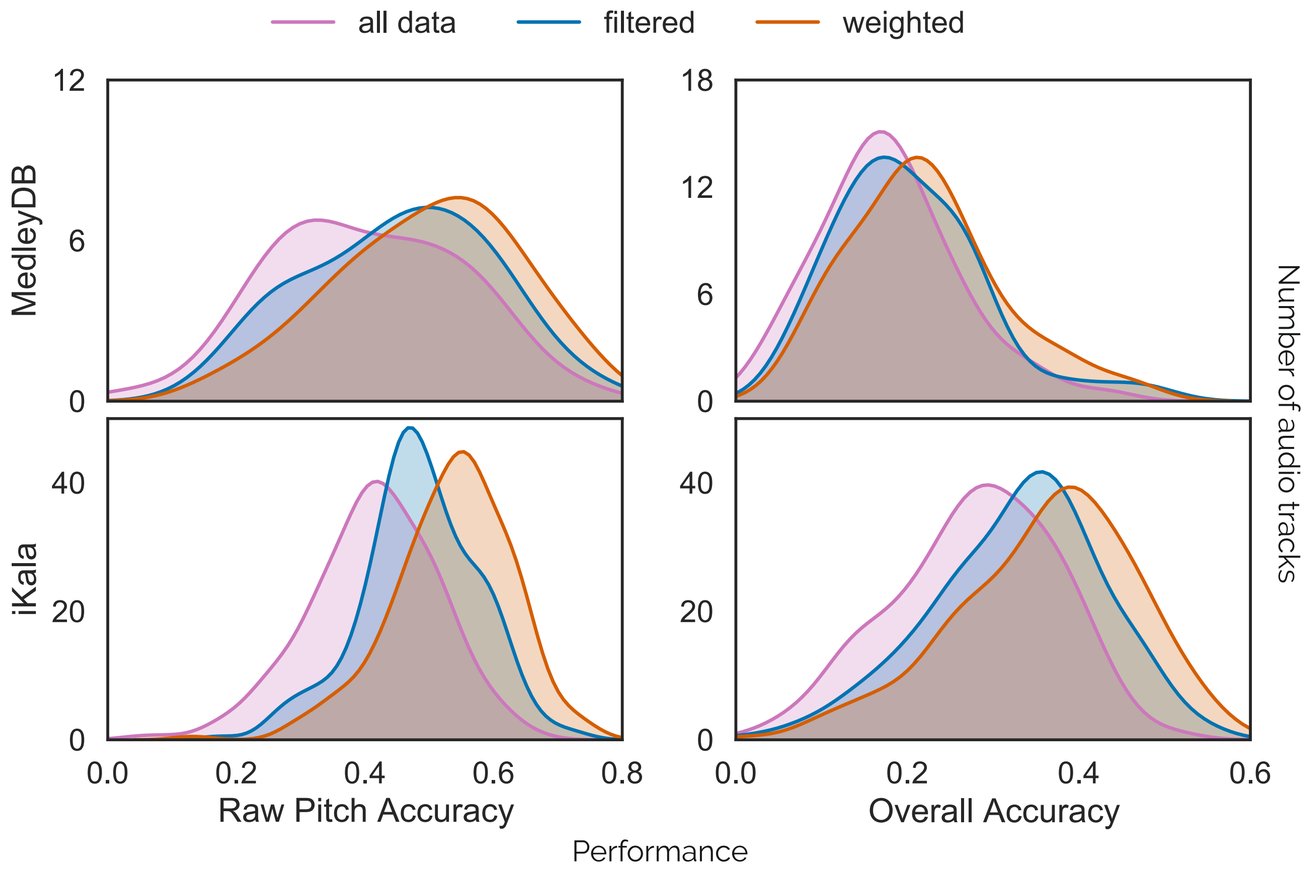
2.3

Training with noisy data

We test the performance of the different models on ground-truths datasets:

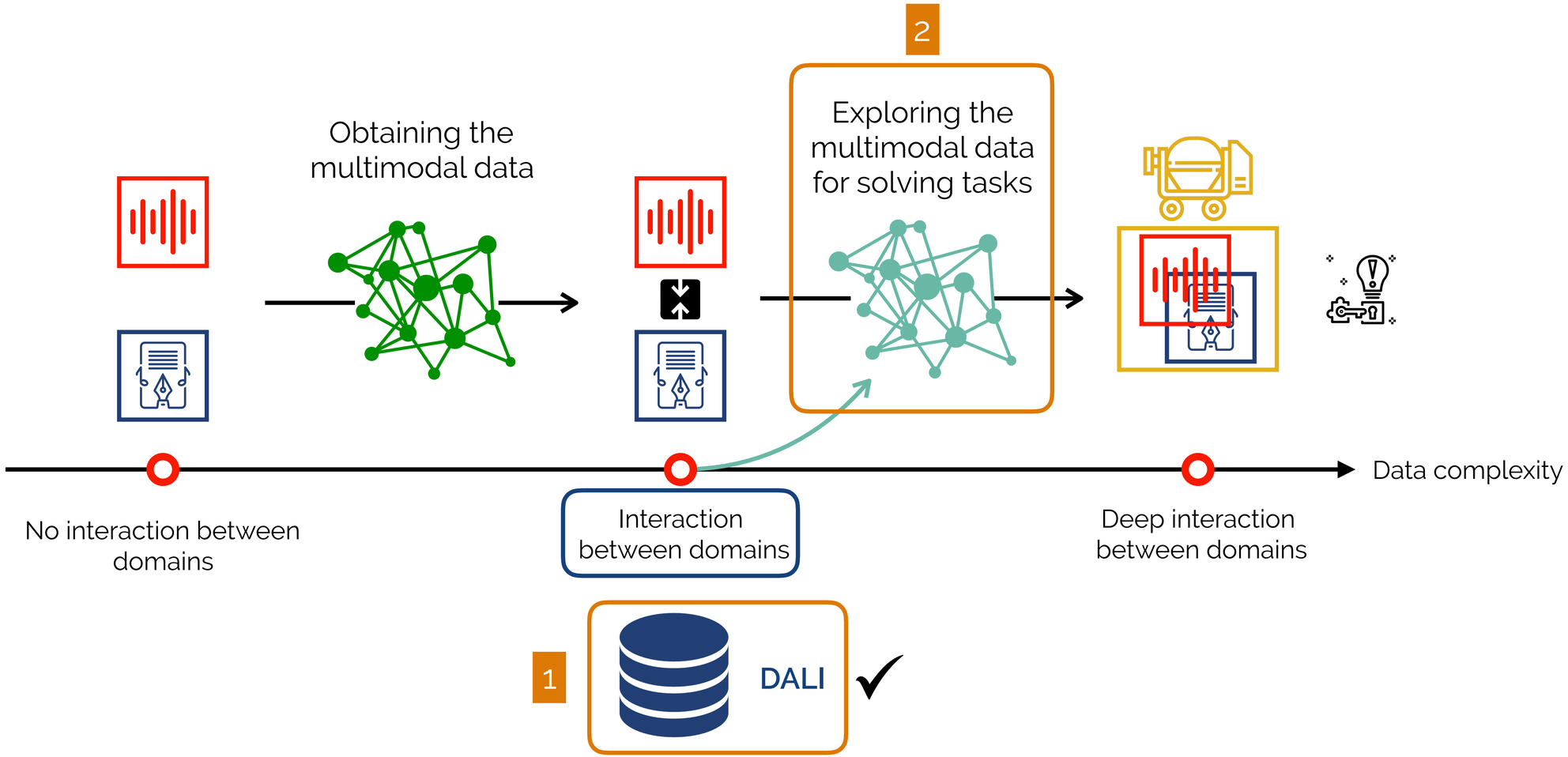
Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., & Bello, J. P. (2014) **Medleydb: A multitrack dataset for annotation-intensive mir research.** (ISMIR)

Chan, T. S., Yeh, T. C., Fan, Z. C., Chen, H. W., Su, L., Yang, Y. H., & Jang, R. (2015) **Vocal activity informed singing voice separation with the iKala dataset.** (ICASSP)



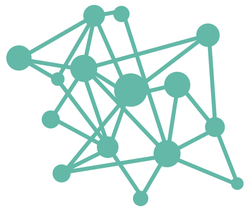
Gabriel Meseguer-Brocal, Rachel Bittner, Simon Duran, Brian Brost. (2020) **Self-Supervised Data Cleansing for Vocal Note Event Annotations.** (ISMIR) Under review

Summary


2
Exploring the multimodal data for solving tasks
Obtaining the multimodal data
Data complexity
No interaction between domains
Interaction between domains
Deep interaction between domains
1
DALI

Plan

1. Introduction
2. Dataset of Aligned Lyric Information - DALI
   - 2.1. Motivation
   - 2.2. Creation
   - 2.3. Training with noisy data
3. Multimodal tasks
   - 3.1. Structures analysis
   - 3.2. Source separations
     - 3.2.1.1. Multitasks
     - 3.2.1.2. Vocals
4. Conclusions and future work



Structure Analysis

3.1

Lyrics segmentation

First multimodal case scenario:

I love you
is all that you can say
words dont come easily
like I love you I love you
but you can say baby
baby can I hold you tonight
baby if I told you the right words
oooh at the right time
you'll be mine

I love you	y=0
is all that you can say	y=0
words dont come easily	y=0
like I love you I love you	y=1
but you can say baby	y=0
baby can I hold you tonight	y=0
baby if I told you the right words	y=0
oooh at the right time	y=1
you'll be mine	

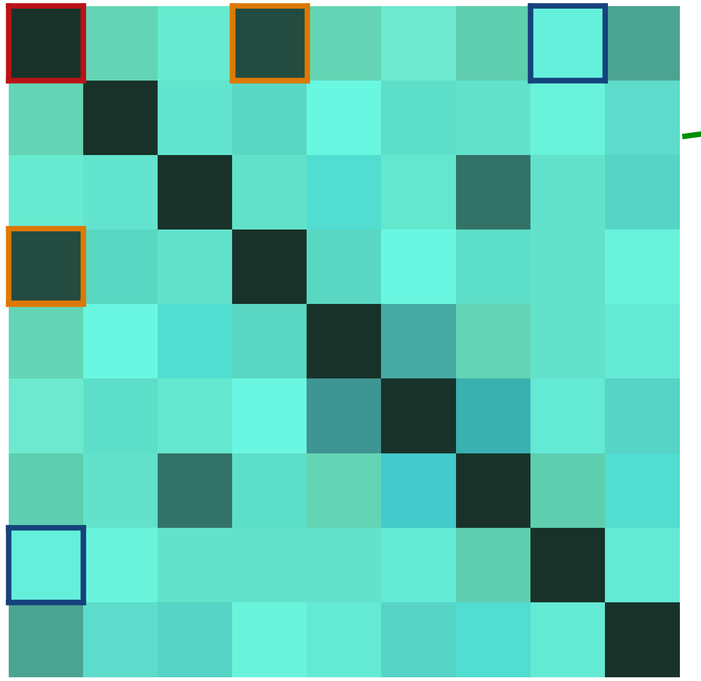
3.1

Lyrics segmentation

Similarity between text lines
using a edit-distance

I love you
is all that you can say
words dont come easily
like I love you I love you
but you can say baby
baby can I hold you tonight
baby if I told you the right words
oooh at the right time
you'll be mine

Self similarity matrix



activation

- +

Fell, Michael and Nechaev, Yaroslav and Cabrio, Elena and Gandon, Fabien. (2018).
Lyrics Segmentation: Textual Macrostructure Detection using Convolutions. In COLING 2018.

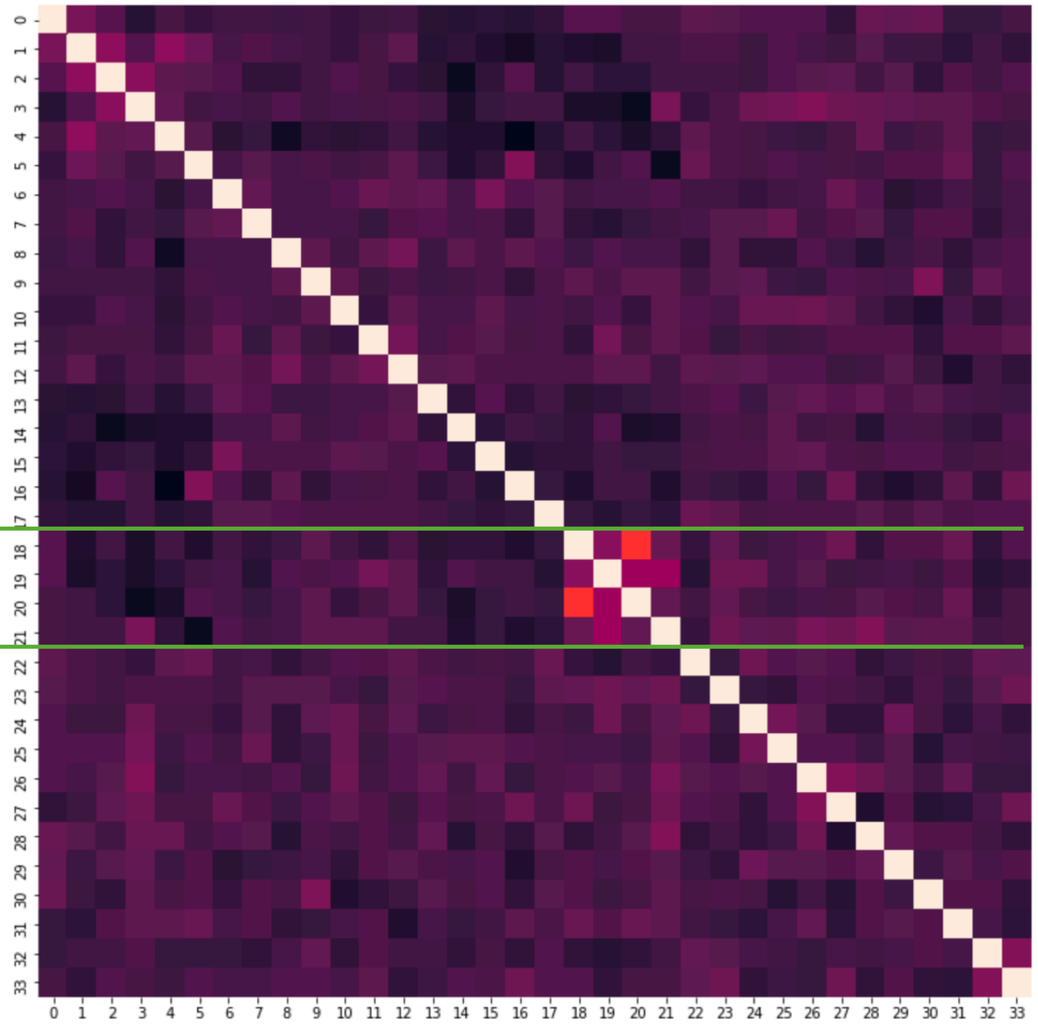
3.1

Lyrics segmentation

```
 0    My click picks your pockets,
 1    splits tha profits,
 2    choose the targets,
 3    and take out the garbage,
 4    Fill tha cartridge,
 5    more lead than Zepplin,
 6    tryin not to step in the blood I left em' in.
 7    Wrestlin with withdrawls got me sippin purple medicine,
 8    Dead rpesidents movin up like Jeffersons.
 9    Lets begin, it began, so begun
10    The story of one who went from girl to woman.
11    From crums to slums, to laced in diamonds.
12    Son by guns that sound like drums.
13    Its no ones fault his life came to a hault, should I open
14    opened up my sliss more.
15    But instead h chose to be another skeleton, I popped his
16    looked like gelatin.
17    A defenition of a cursed individual, original, you caint

18    1-2 Better call yo crews
19    3-4 Need to lock yo doors
20    5-6 Better loaad yo clips
21    7-8 Time to meet yo fate

22    Some of yall ****  aint wit this,
23    Well then get tha fuck on bout yo business.
24    Cuz my kind dont take kinds who doze,
25    who act like hoes while I sip Irish Rose,
26    And smoke some of the worst weed youd ever tasted,
27    Fuck it blaze it lets all get wasted.
28    Still a Ja-ca, you need to face it,
29    Mad cuz yo house costs as much as yo bracelet,
30    Never patient, compitition, sound ancient,
31    Im gettin followed by a Federal Agent,
32    Engagements, gangsta banquets, piggarements, stainless,
33    Leave you brainless make it painless.
```



3.1

Lyrics segmentation

Hypothesis: since melodies are often repeated, the part of the structure which is not capture in the text may arise from the audio.

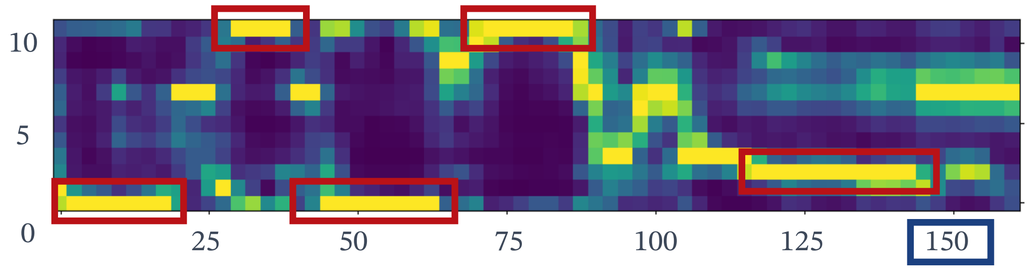
In ***DALI***:

How to measure **audio** similarity between two segments of **different duration**?

3.1

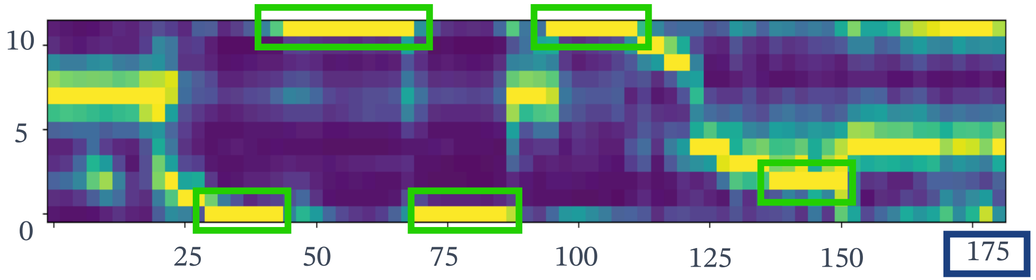
Lyrics segmentation

s_0 with duration d_0

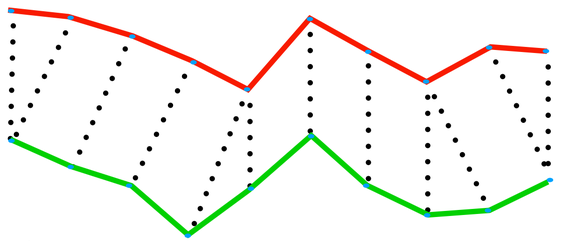


Feature extractor that highlight a musical dimension

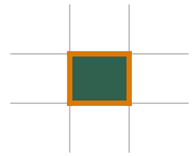
s_1 with duration d_1



Dynamic Time Warping



=



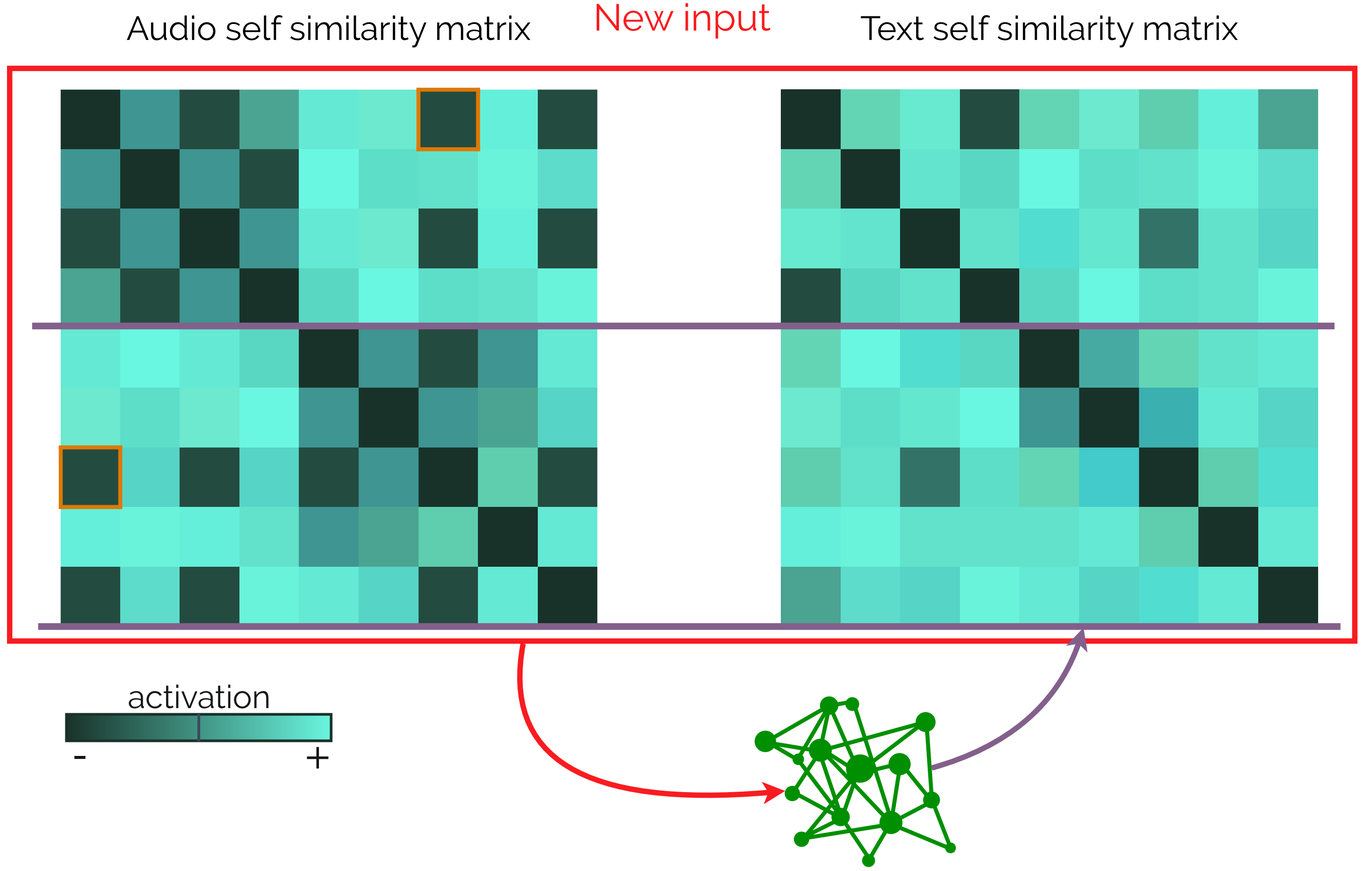
Cell in the music self similarity matrix

3.1

Lyrics segmentation

Hypothesis: since melodies are often repeated, the part of the structure which is not capture in the text may arise from the audio.

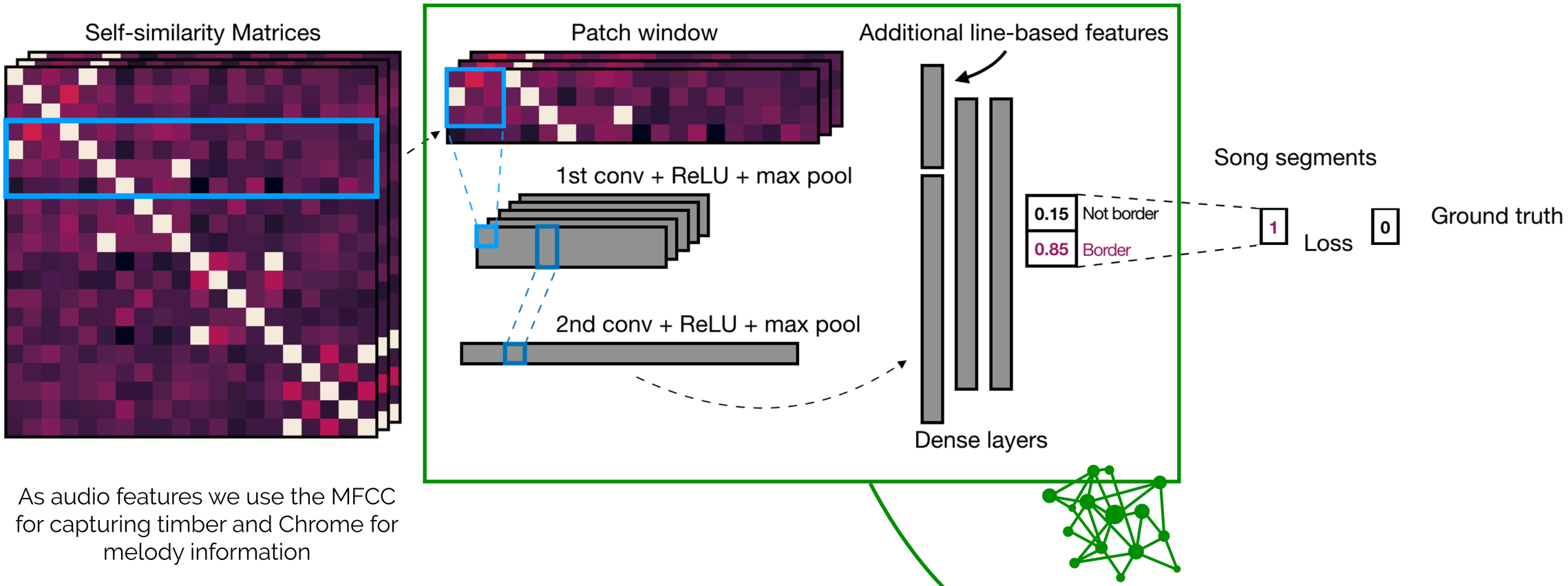
In ***DALI***:



Michael Fell E, Yaroslav Nechaev, Gabriel Meseguer-Brocal, Elena Cabrio, Fabien Gandon and Geoffroy Peeters (2020). ***Lyrics Segmentation via Bimodal Text-audio Representation***. Natural Language Engineering. Under review

3.1

Lyrics segmentation



As audio features we use the MFCC for capturing timber and Chrome for melody information

Cohen-Hadria, A., & Peeters, G. (2017). ***Music structure boundaries estimation using multiple self-similarity matrices as input depth of convolutional neural networks***. *AES International Conference on Semantic Audio*. Audio Engineering Society.

Fell, Michael and Nechaev, Yaroslav and Cabrio, Elena and Gandon, Fabien. (2018). ***Lyrics Segmentation: Textual Macrostructure Detection using Convolutions***. In COLING 2018.

Lyrics segmentation

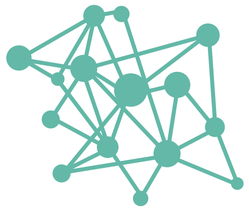
Dataset	Model	Features	P	R	F_1
M^+	*text*	{str}	78.7	64.2	70.8
	audio	{mfcc, chroma}	79.2	63.8	70.4
	multi	{str, mfcc, chroma}	82.7	70.3	75.3
M^0	*text*	{str}	73.6	54.5	62.8
	audio	{mfcc, chroma}	74.9	48.9	59.5
	multi	{str, mfcc, chroma}	75.8	59.4	66.5
M^-	*text*	{str}	67.5	30.9	41.9
	audio	{mfcc, chroma}	66.1	24.7	36.1
	multi	{str, mfcc, chroma}	68.0	35.8	46.7

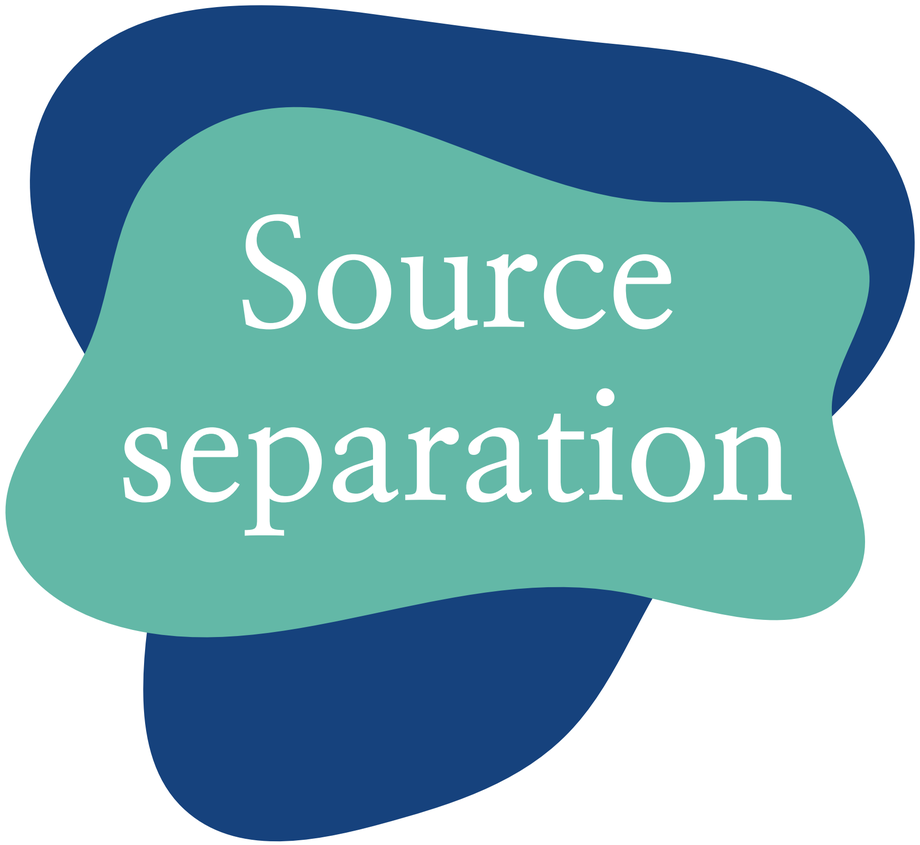
Different versions of DALI according to a confident measurement

Michael Fell E, Yaroslav Nechaev, Gabriel Meseguer-Brocal, Elena Cabrio, Fabien Gandon and Geoffroy Peeters (2020). ***Lyrics Segmentation via Bimodal Text-audio Representation***. Natural Language Engineering. Under review

Plan

1. Introduction
2. Dataset of Aligned Lyric Information - DALI
   - 2.1. Motivation
   - 2.2. Creation
   - 2.3. Training with noisy data
3. Multimodal tasks
   - 3.1. Structures analysis
   - 3.2. Source separations
     - 3.2.1.1. Multitasks
     - 3.2.1.2. Vocals
4. Conclusions and future work

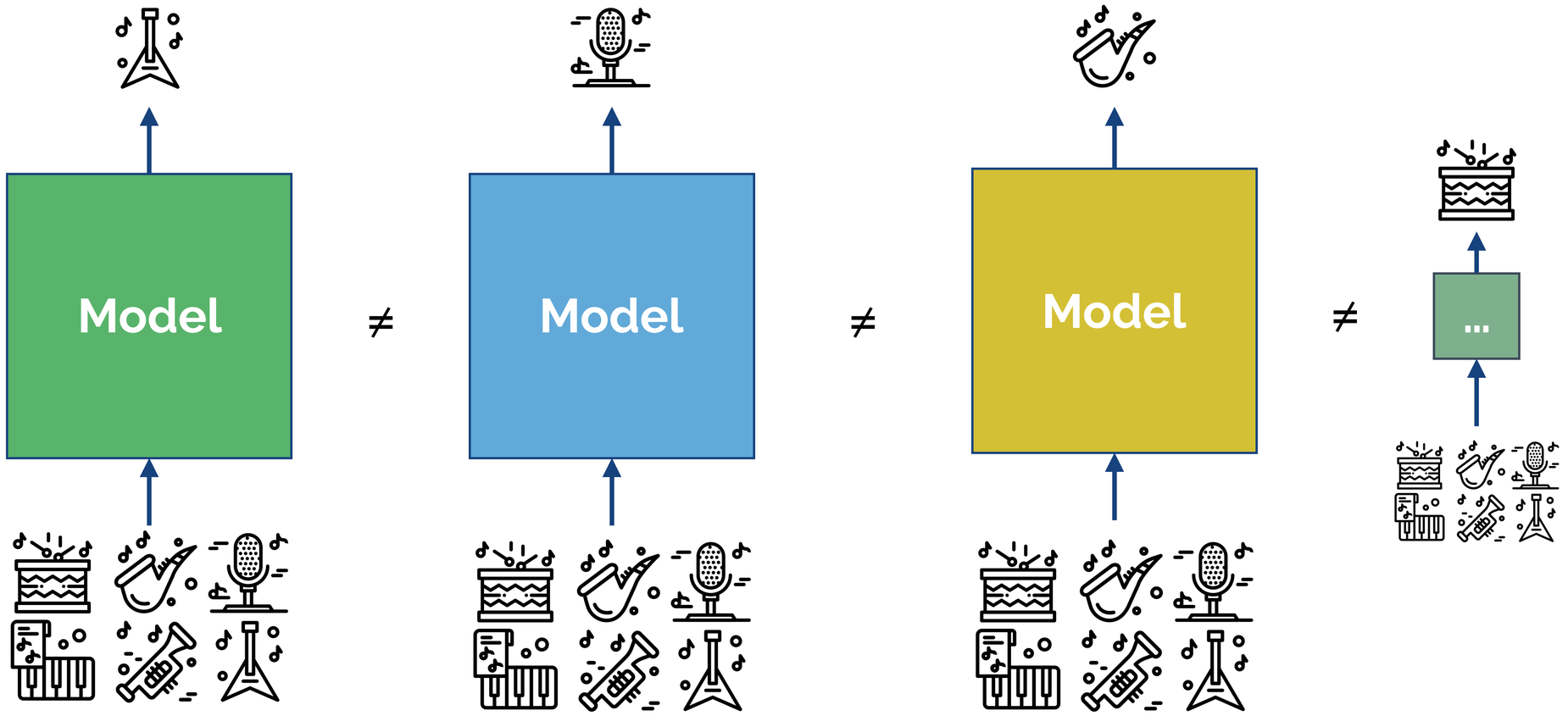



Source
separation

3.2

Source separation

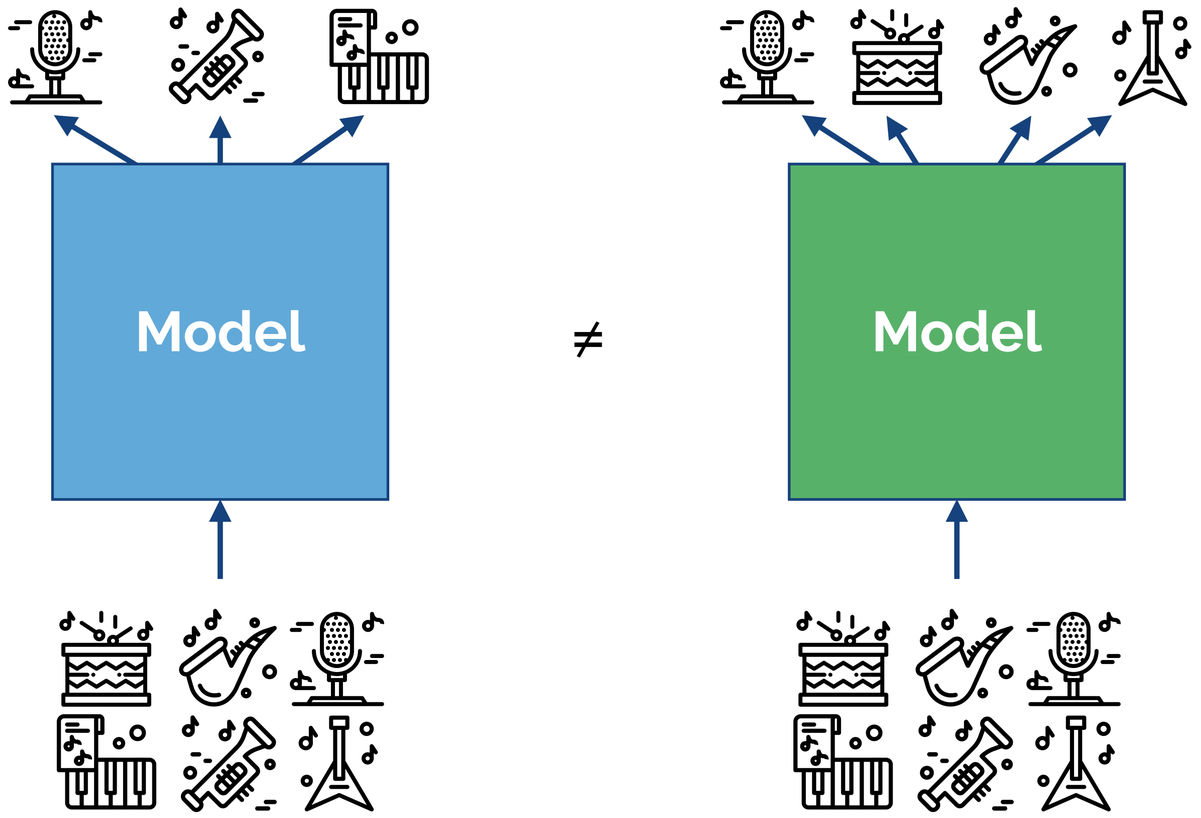
A dedicated model trained for separating each instrument.



3.2

Source separation

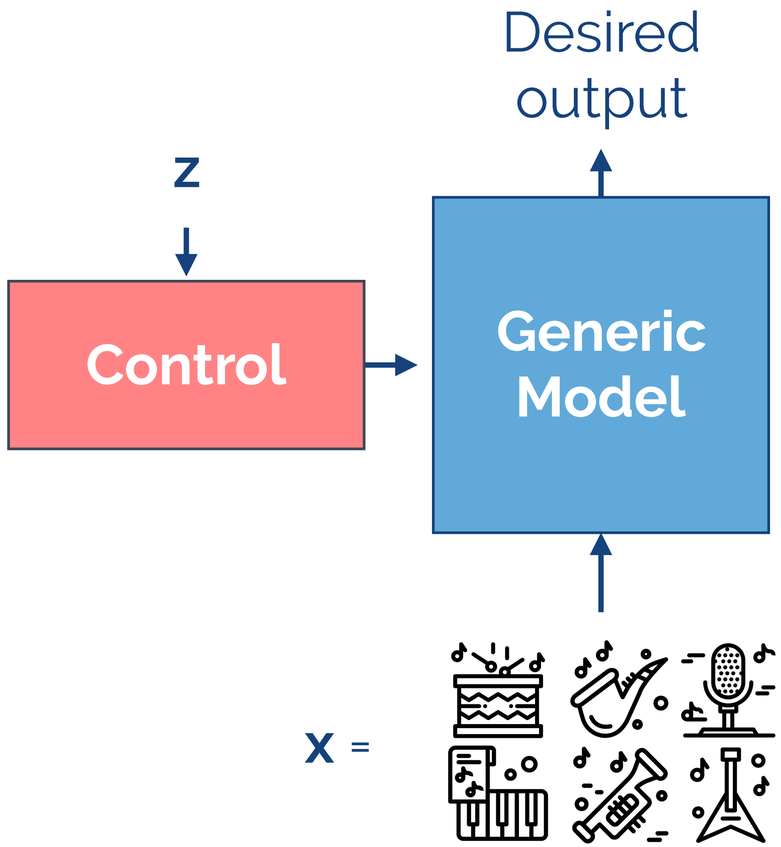
A single model with a **fixed** number of instrument separations.



3.2

Source separation

Core idea: An input ***x*** is processed differently depending on external context **z**

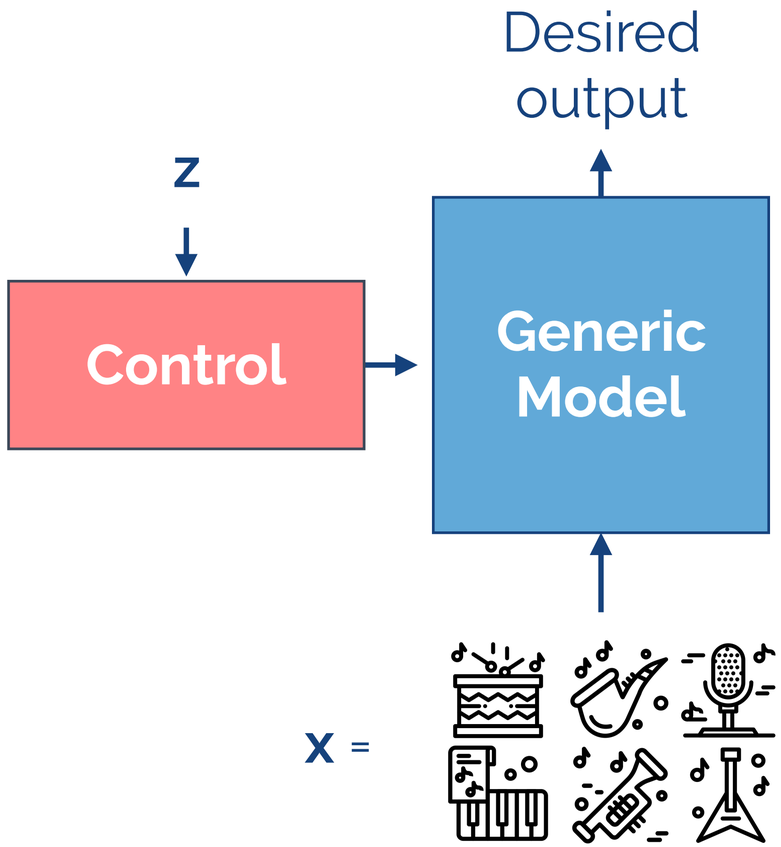


Why?

- These tasks share many many similarities.
- Open the door to universal models.
- Add additional flexibility to deal with complex problems.

Source separation

Core idea: An input ***x*** is processed differently depending on external context **z**

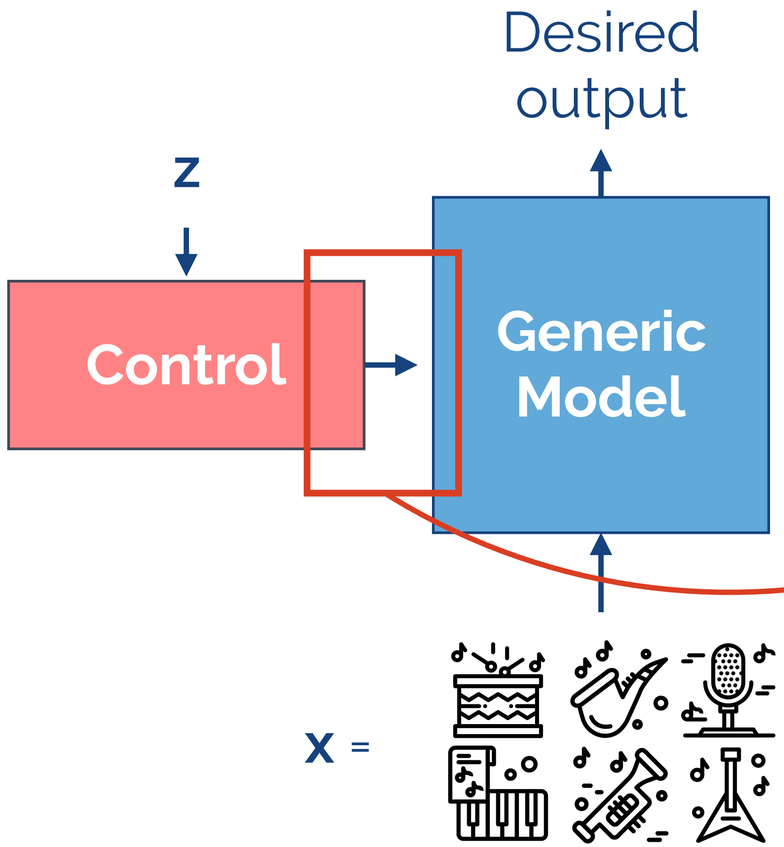


What external information?

1. Instrument id: Multitask sources separation.
2. Phoneme information: perform different source separation operations for each phoneme.

Source separation

Core idea: An input ***x*** is processed differently depending on external context **z**



How?

The control is done via FiLM (Feature-wise transformations) layers

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, Aaron Courville (2017).
FiLM: Visual Reasoning with a General Conditioning Layer
CVPR17

3.2

Source separation

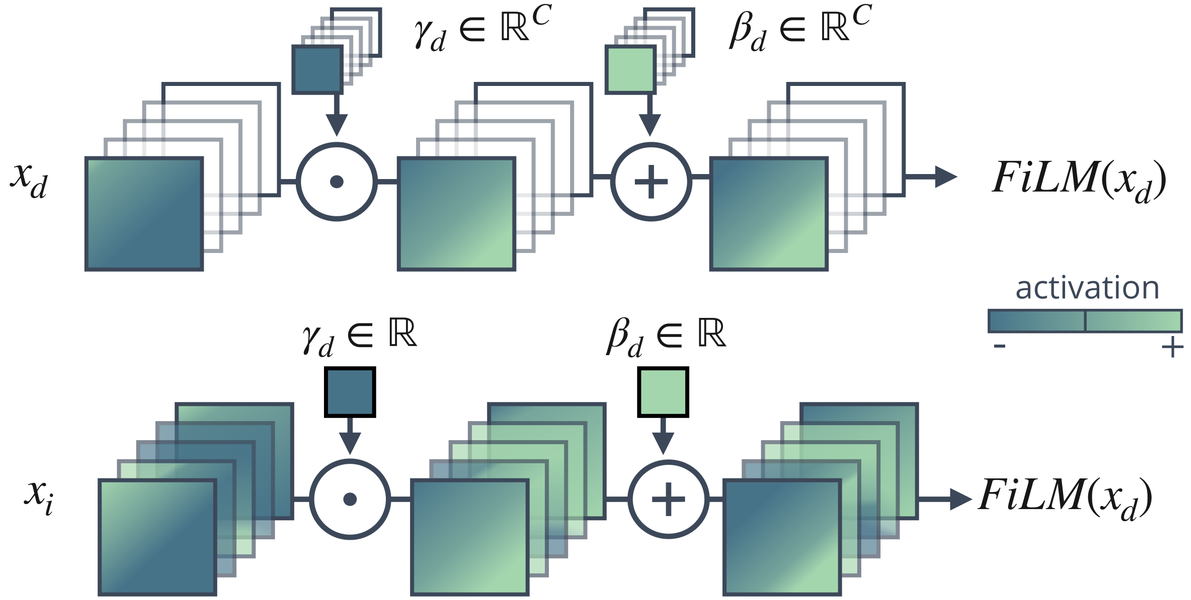
FiLM permits to modulate any neural network architecture inserting one or several FiLM layers at any depth of the original model:

$$FiLM(x_d) = \gamma_d(z) \odot x_d + \beta(z)_d$$

$\mathbf{x}$ is the input of the FiLM layer and γ and β the learnable parameters that scale and shift $\mathbf{x}$ based on an external information, $\mathbf{z}$.

FiLM complex layer *(Co)*: independent affine transformations are applied to each feature map $\boldsymbol{c}$.

FiLM simple layer *(Si)*: same affine transformation to all the input feature maps $\boldsymbol{x}$.

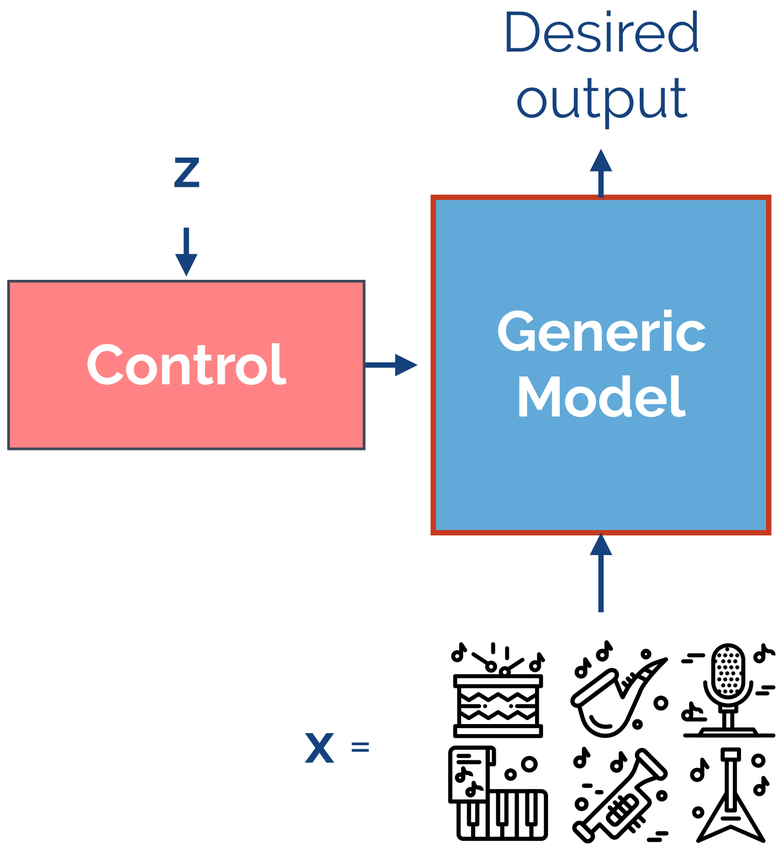


Gabriel Meseguer-Brocal and Geoffroy Peeters (2019). **Conditioned-U-Net: Introducing a Control Mechanism in the U-Net for Multiple Source Separations** (ISMIR)

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, Aaron Courville (2017). **FiLM: Visual Reasoning with a General Conditioning Layer** CVPR07

Source separation

Core idea: An input ***x*** is processed differently depending on external context **z**

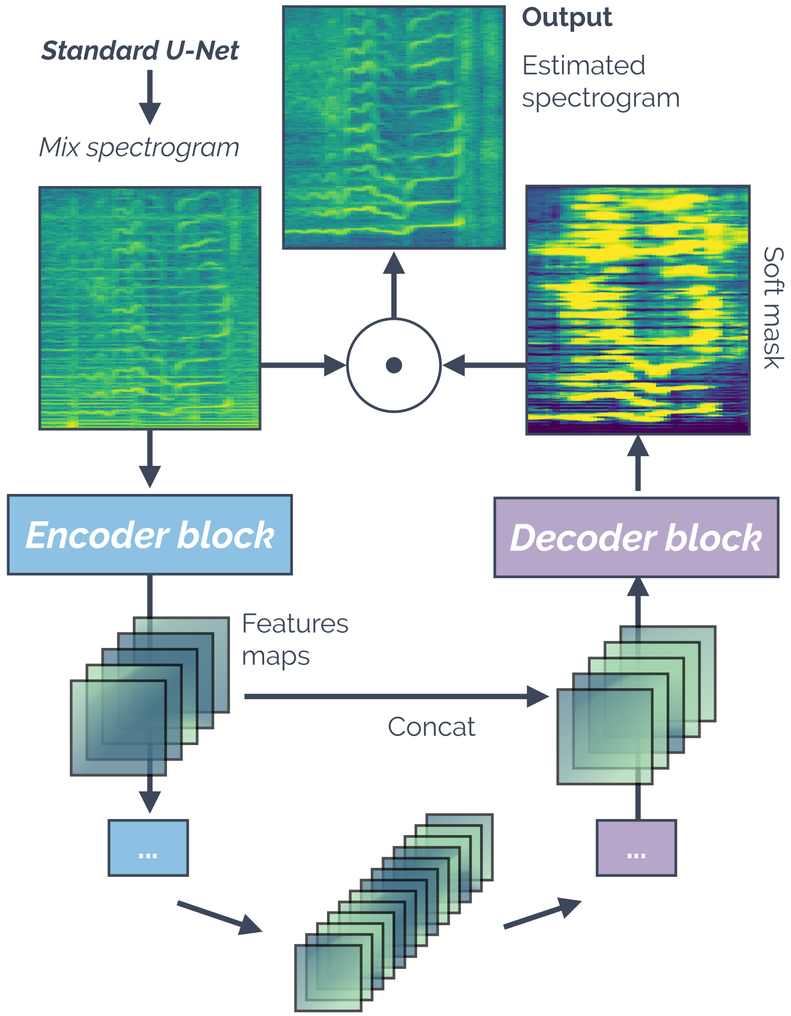


Source separation

U-NET

Generic Model

=

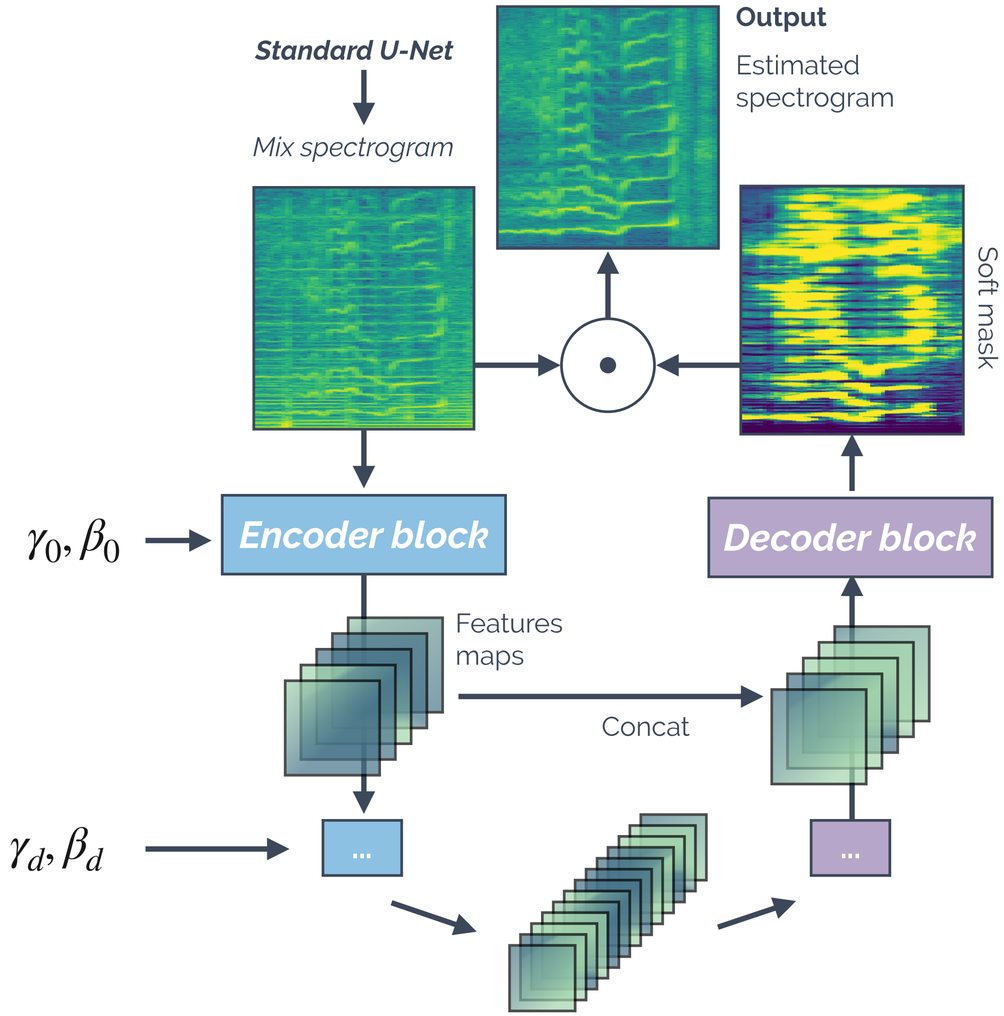


When?

- **Encoder**: codifies and highlights the relevant information from the signal to separate a particular instrument creating a latent space
- **Decoder**: transforms the latent space back to audio signal.

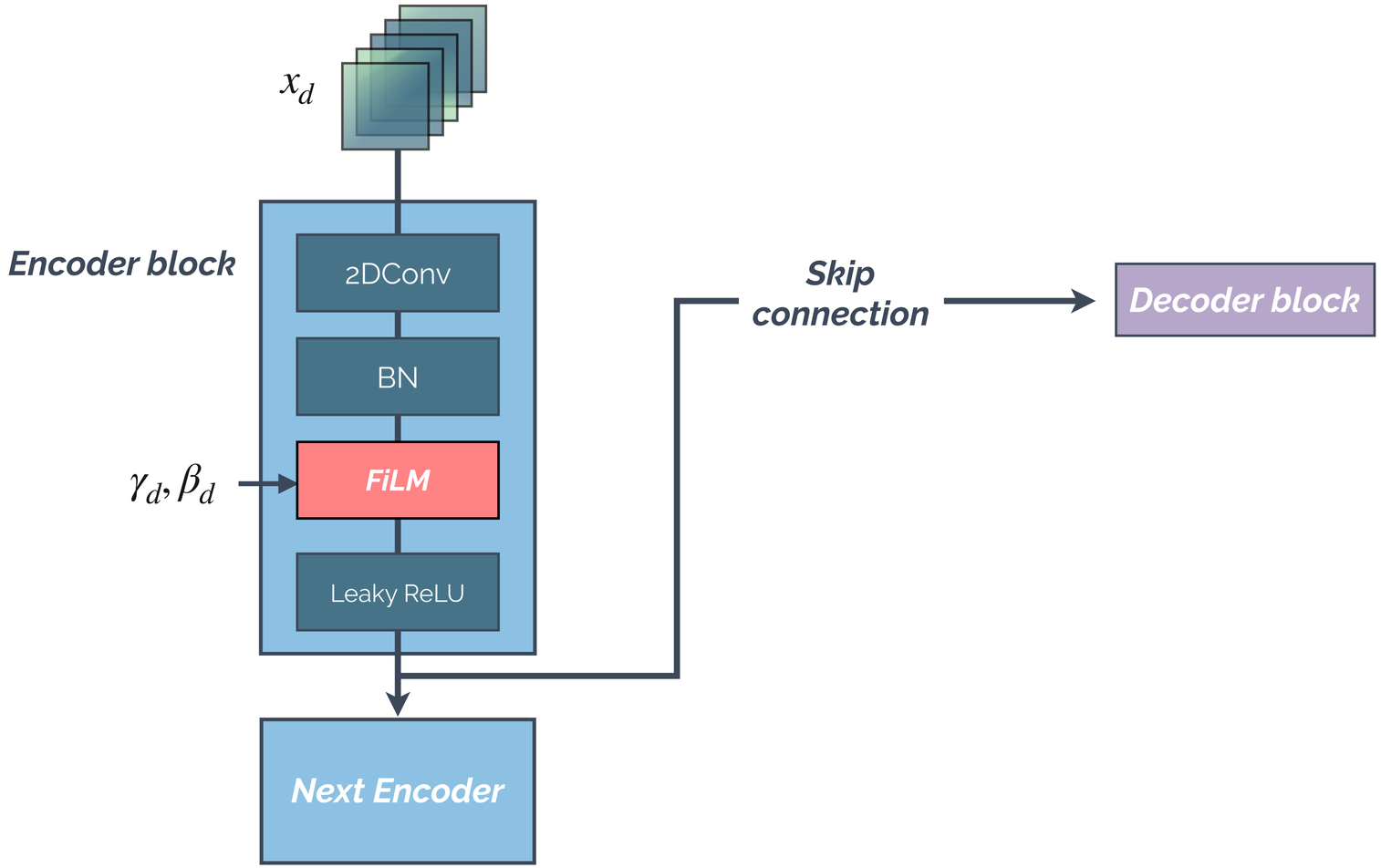
Jansson, Andreas et all. (2017). **Singing voice separation with deep u-net convolutional networks** (ISMIR)

Source separation



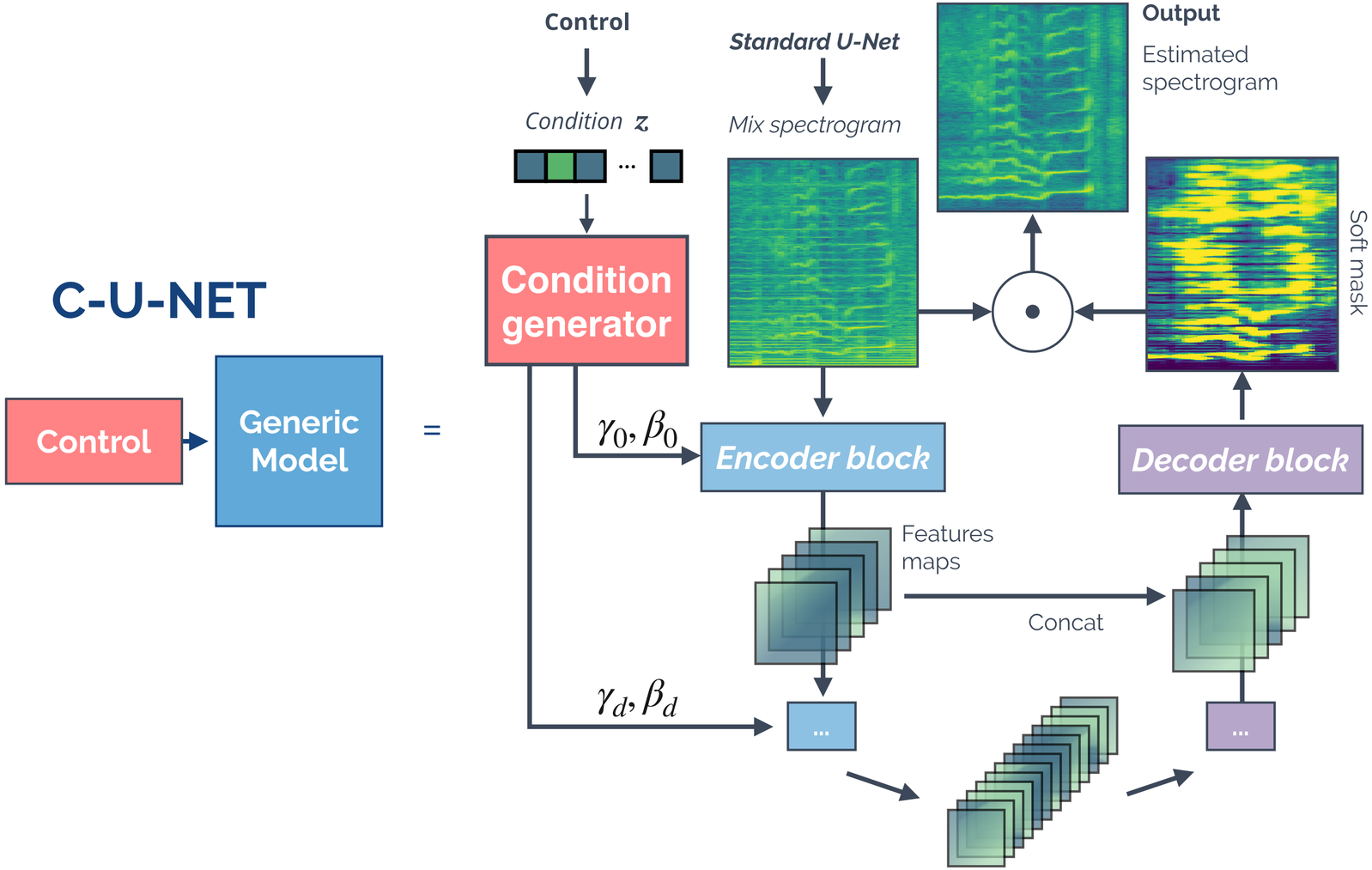
Gabriel Meseguer-Brocal and Geoffroy Peeters (2019). **Conditioned-U-Net: Introducing a Control Mechanism in the U-Net for Multiple Source Separations** (ISMIR)

Source separation



Gabriel Meseguer-Brocal and Geoffroy Peeters (2019). **Conditioned-U-Net: Introducing a Control Mechanism in the U-Net for Multiple Source Separations** (ISMIR)

Source separation



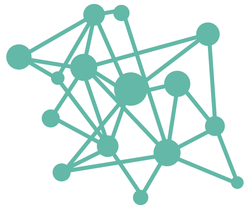
What external information?

1. Instrument id: Multitask sources separation.
2. Phoneme information: perform different source separation operations for each phoneme.

Gabriel Meseguer-Brocal and Geoffroy Peeters (2019). **Conditioned-U-Net: Introducing a Control Mechanism in the U-Net for Multiple Source Separations** (ISMIR)

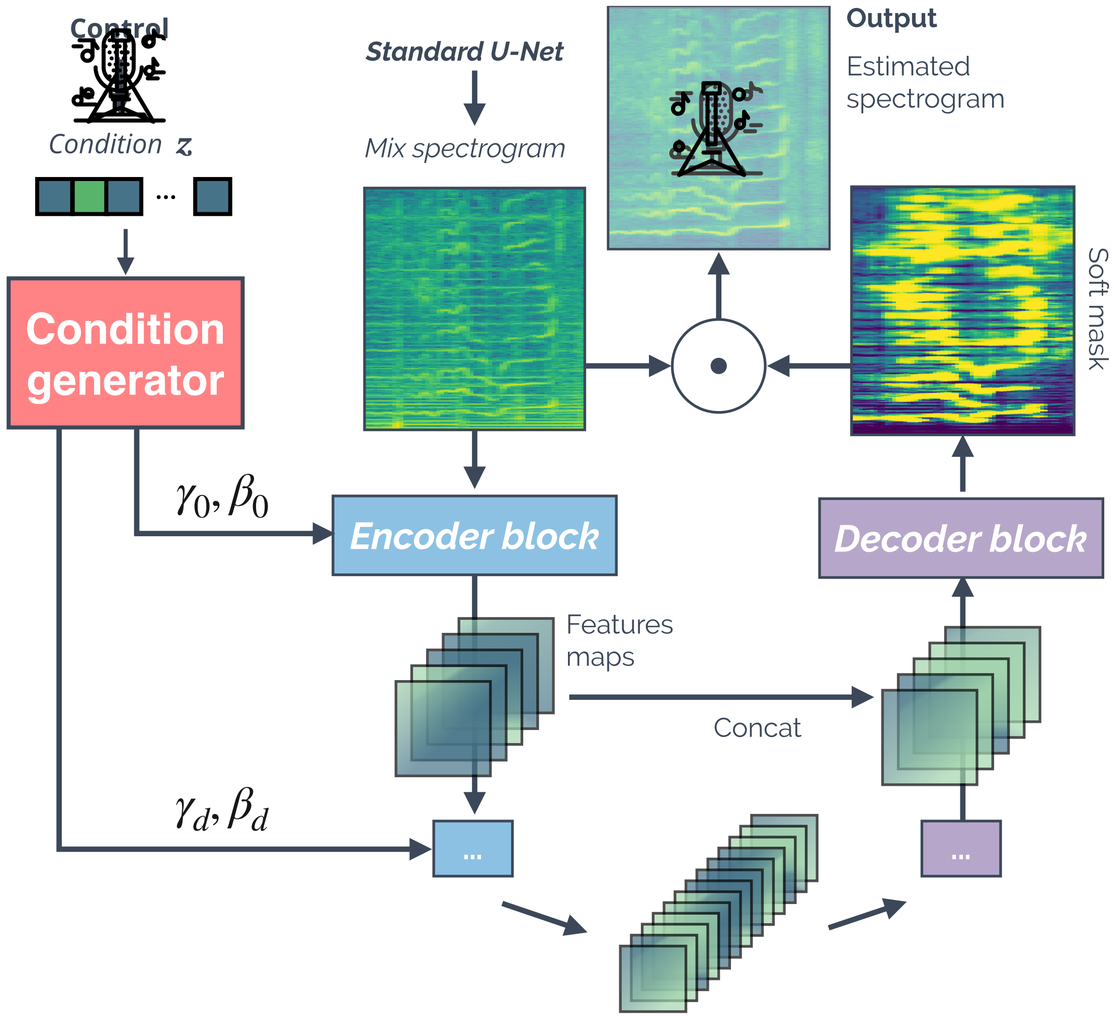
Plan

1. Introduction
2. Dataset of Aligned Lyric Information - DALI
   - 2.1. Motivation
   - 2.2. Creation
   - 2.3. Training with noisy data
3. Multimodal tasks
   - 3.1. Structures analysis
   - 3.2. Source separations
     - → 3.2.1.1. Multitasks
     - 3.2.1.2. Vocals
4. Conclusions and future work



3.2.1

Source separation – Multitask



Rafii, Zafar and Liutkus, Antoine and Fabian-Robert Stöter and Mimilakis, Stylianos Ioannis and Bittner, Rachel (2017) **The MUSDB18 corpus for music separation**

What external information?

1. Instrument id: Multitask sources separation.
2. Phoneme information: perform different source separation operations for each phoneme.

Gabriel Meseguer-Brocal and Geoffroy Peeters (2019). **Conditioned-U-Net: Introducing a Control Mechanism in the U-Net for Multiple Source Separations** (ISMIR)

Source separation – Multitask

MODEL	Total		
	SIR	SAR	SDR
Fix-U-Net(x4)	7.31 ± 4.04	5.70 ± 3.10	2.36 ± 3.96
C-U-Net-SiC-np	7.35 ± 4.13	5.74 ± 3.18	2.34 ± 3.69
C-U-Net-SiC-p	**8.00** ± 4.37	**5.74** ± 3.63	**2.54** ± 4.07
C-U-Net-CoC-np	7.27 ± 4.24	5.60 ± 2.88	2.36 ± 3.81
C-U-Net-CoC-p	7.49 ± 4.54	5.67 ± 3.03	2.42 ± 4.21
C-U-Net-SiF-np	7.23 ± 3.97	5.59 ± 3.01	2.22 ± 3.67
C-U-Net-SiF-p	7.64 ± 4.05	5.73 ± 2.88	2.46 ± 3.88
C-U-Net-CoF-np	7.42 ± 4.20	5.59 ± 3.07	2.32 ± 3.85
C-U-Net-CoF-p	7.52 ± 4.04	5.71 ± 2.99	2.42 ± 3.97

	Model	SIR	SAR	SDR
Vocals	*Fix-U-Net(x4)*	10.70 ± 4.26	5.39 ± 3.58	3.52 ± 4.88 (4.72)
	C-U-Net-CoF	10.76 ± 4.39	5.32 ± 3.27	3.50 ± 4.37 (4.65)
	Wave-U-Net-D	-	-	0.55 ± 13.67 (4.58)
	Wave-U-Net-M	-	-	-2.10 ± 15.41 (3.0)
Drums	*Fix-U-Net(x4)*	10.08 ± 4.28	6.42 ± 3.28	4.28 ± 3.65 (4.13)
	C-U-Net-CoF	10.03 ± 4.34	6.80 ± 3.25	4.30 ± 3.81 (4.38)
	Wave-U-Net-M	-	-	2.88 ± 7.68 (4.15)
Bass	*Fix-U-Net(x4)*	4.64 ± 4.76	6.51 ± 2.68	1.46 ± 4.31 (2.48)
	C-U-Net-CoF	5.30 ± 4.73	6.29 ± 2.39	1.65 ± 4.07 (2.60)
	Wave-U-Net-M	-	-	-0.30 ± 13.50 (2.91)
Rest	*Fix-U-Net(x4)*	3.83 ± 2.84	4.47 ± 2.85	0.19 ± 3.00 (0.97)
	C-U-Net-CoF	4.00 ± 2.70	4.37 ± 3.06	0.24 ± 3.64 (1.71)
	Wave-U-Net-M	-	-	1.68 ± 6.14 (2.03)

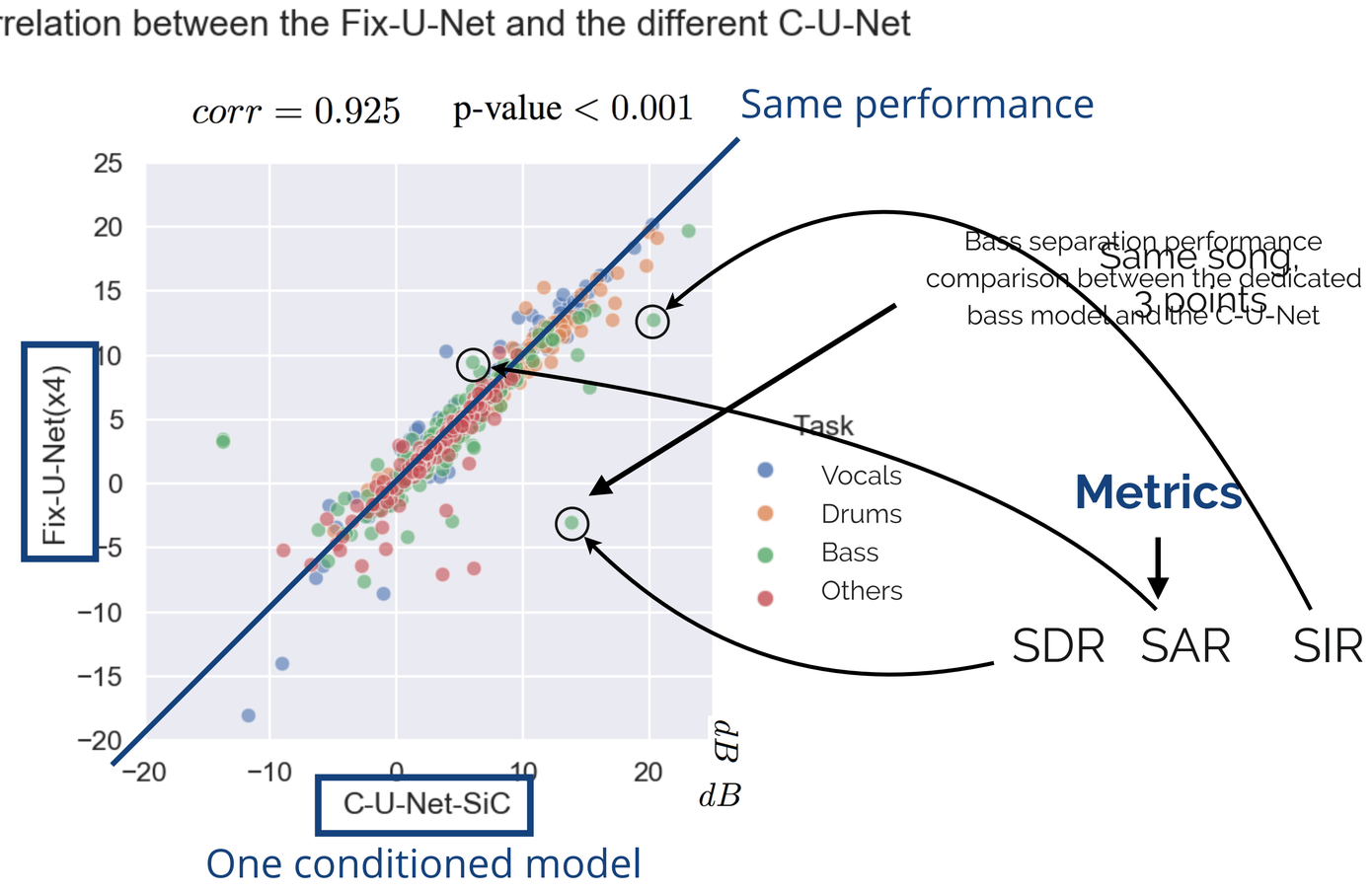
The C-U-Net perform similarly to the four dedicate U-Nets.

Gabriel Meseguer-Brocal and Geoffroy Peeters (2019). **Conditioned-U-Net: Introducing a Control Mechanism in the U-Net for Multiple Source Separations** (ISMIR)

3.2.1

Source separation – Multitask

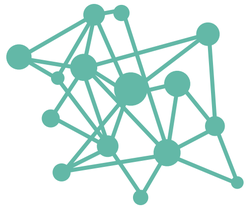
4 dedicated models trained for separation each instrument.



One conditioned model

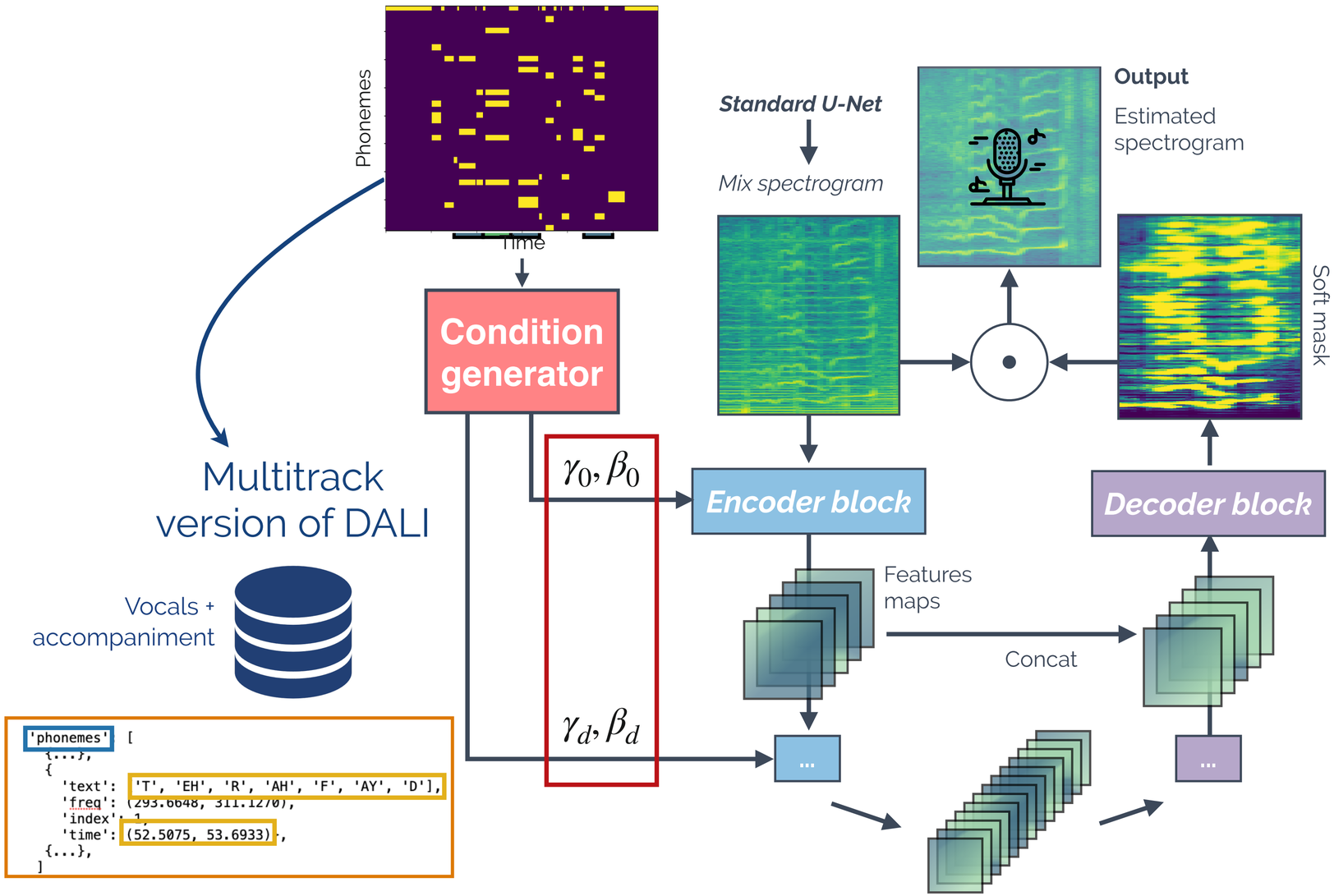
Plan

1. Introduction
2. Dataset of Aligned Lyric Information - DALI
   - 2.1. Motivation
   - 2.2. Creation
   - 2.3. Training with noisy data
3. Multimodal tasks
   - 3.1. Structures analysis
   - 3.2. Source separations
     - 3.2.1.1. Multitasks
     - → 3.2.1.2. Vocals
4. Conclusions and future work



3.2.2

Source separation – Vocals



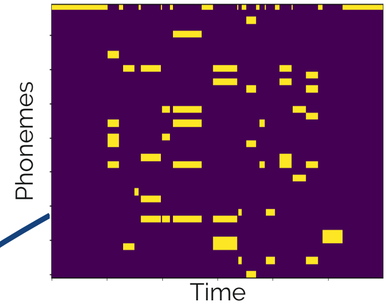
What external information?

1. Instrument id: Multitask sources separation.
2. Phoneme information: perform different source separation operations for each phoneme.

Gabriel Meseguer-Brocal and Geoffroy Peeters (2020). **Content based singing voice source separation via strong conditioning using aligned phonemes** (ISMIR) Under review

3.2.2

Source separation – Vocals



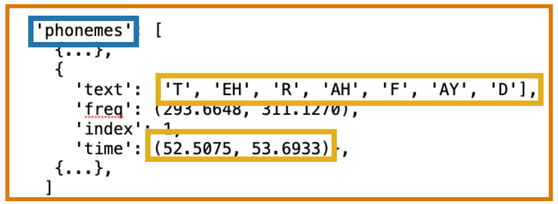
Condition generator

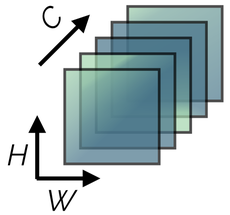
γ_0, β_0

γ_d, β_d

Multitrack version of DALI

Vocals + accompaniment





$x_d \in \mathbb{R}^{W \times H \times C}$ "Time" "Freq" Channels

$$FiLM(x_d) = \gamma_d(z) \odot x_d + \beta(z)_d$$

$(\gamma_d, \beta_d) \in \mathbb{R}^C$ Complex FiLM

$(\gamma_d, \beta_d) \in \mathbb{R}$ Simple FiLM

We are loosing the time information!!

Weak conditioning: same FiLM operation for an input patch.

Gabriel Meseguer-Brocal and Geoffroy Peeters (2020). **Content based singing voice source separation via strong conditioning using aligned phonemes** (ISMIR) Under review

3.2.2

Source separation – Vocals

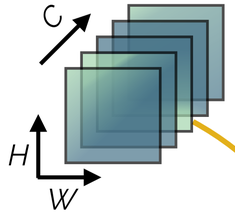
$$FiLM(x_d) = \gamma_d(z) \odot x_d + \beta(z)_d$$

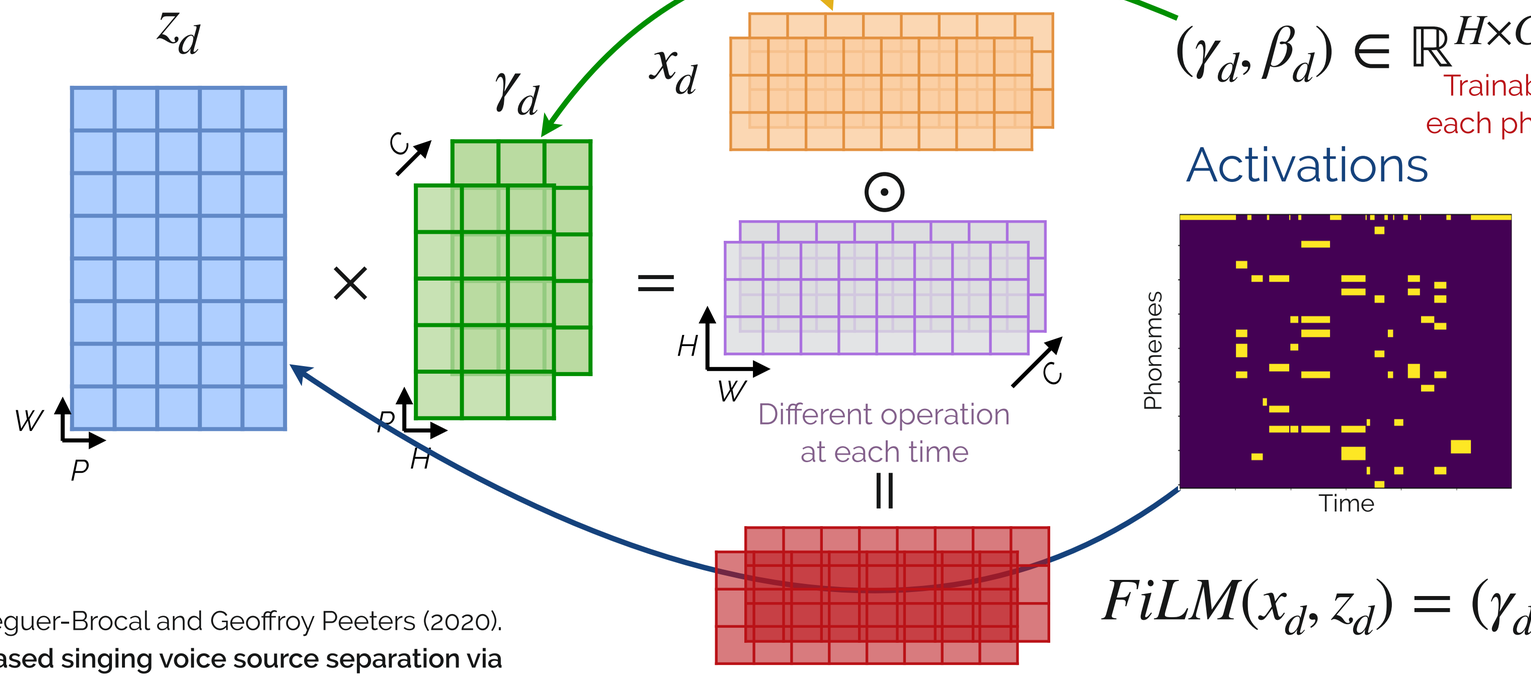
$(\gamma_d, \beta_d) \in \mathbb{R}^C$ Complex FiLM

$(\gamma_d, \beta_d) \in \mathbb{R}$ Simple FiLM

Weak

$x_d \in \mathbb{R}^{W \times H \times C}$ "Time" "Freq" Channels





Basis tensor

$(\gamma_d, \beta_d) \in \mathbb{R}^{H \times C \times P}$

Trainable values for each phonemes class

Activations

Strong: different operations (defined by the activations) at different time.

$$FiLM(x_d, z_d) = (\gamma_d \times z_d) \odot x_d + (\beta_d \times z_d)$$

Gabriel Meseguer-Brocal and Geoffroy Peeters (2020). **Content based singing voice source separation via strong conditioning using aligned phonemes** (ISMIR) Under review

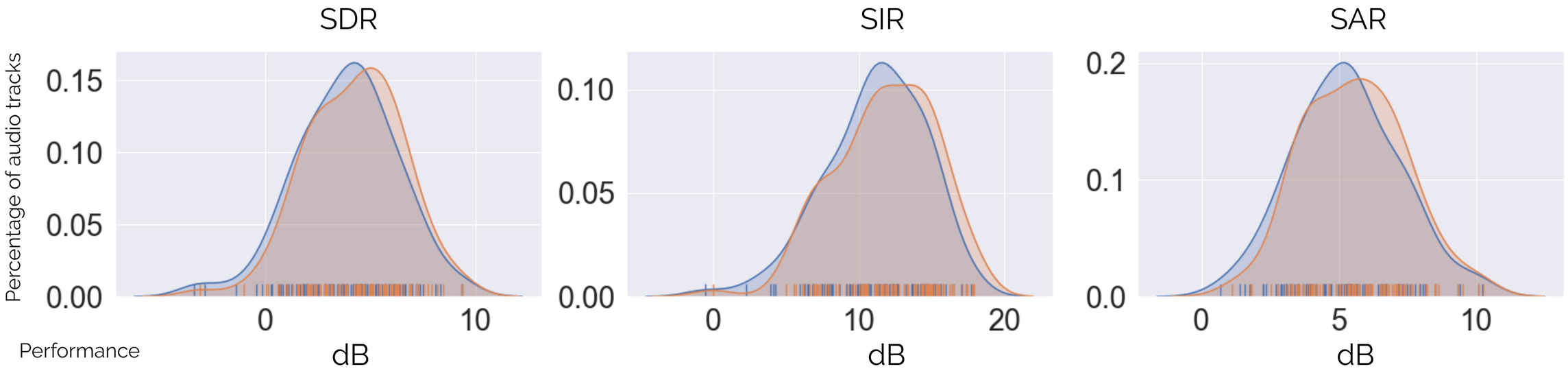
3.2.2

Source separation – Vocals

Original model

Conditioned model

$(\gamma_d, \beta_d) \in \mathbb{R}^P$



Gabriel Meseguer-Brocal and Geoffroy Peeters (2020). **Content based singing voice source separation via strong conditioning using aligned phonemes** (ISMIR) Under review

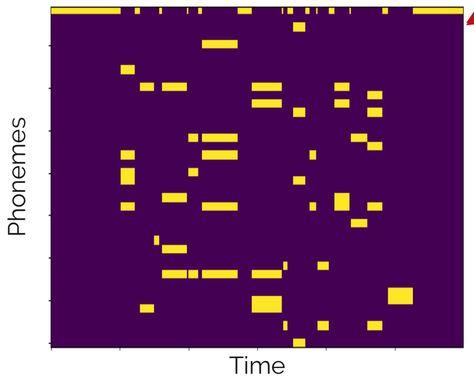
3.2.2

Source separation – Vocals

Limitations:

$$FiLM(x_d, z_d) = (\gamma_d \times z_d) \odot x_d + (\beta_d \times z_d)$$

Activations



Issues:

1. Done by the users = potential bad alignment.
2. Phonemes obtained automatically.
3. Not onset for phonemes inside words.

→ Bad activations hurt the performance

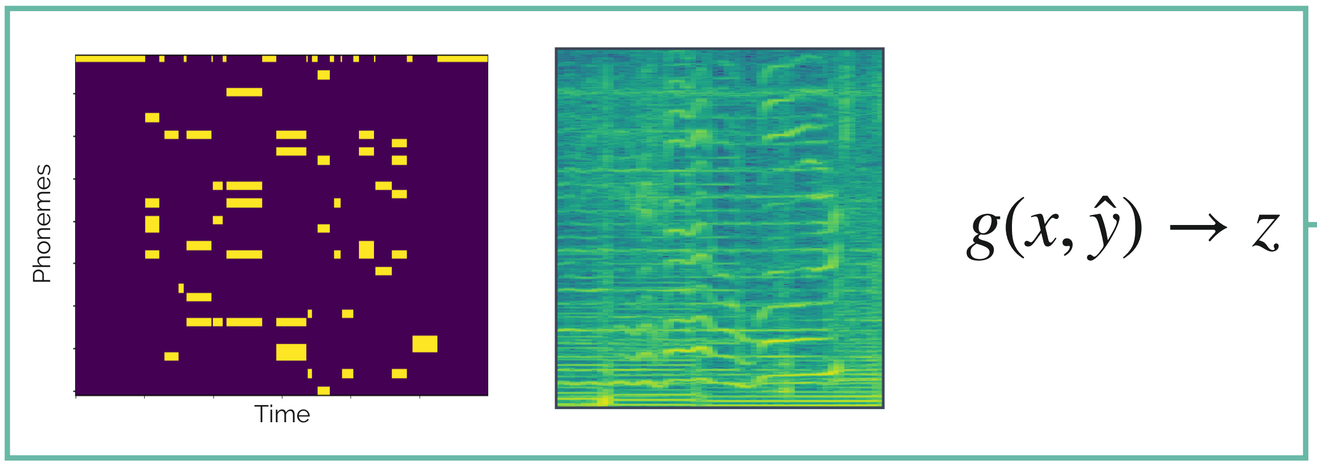
Gabriel Meseguer-Brocal and Geoffroy Peeters (2020). **Content based singing voice source separation via strong conditioning using aligned phonemes** (ISMIR) Under review

3.2.2

Source separation – Vocals

Limitations:

$$FiLM(x_d, z_d) = (\gamma_d \times z_d) \odot x_d + (\beta_d \times z_d)$$

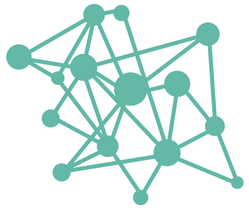


$$g(x, \hat{y}) \rightarrow z$$

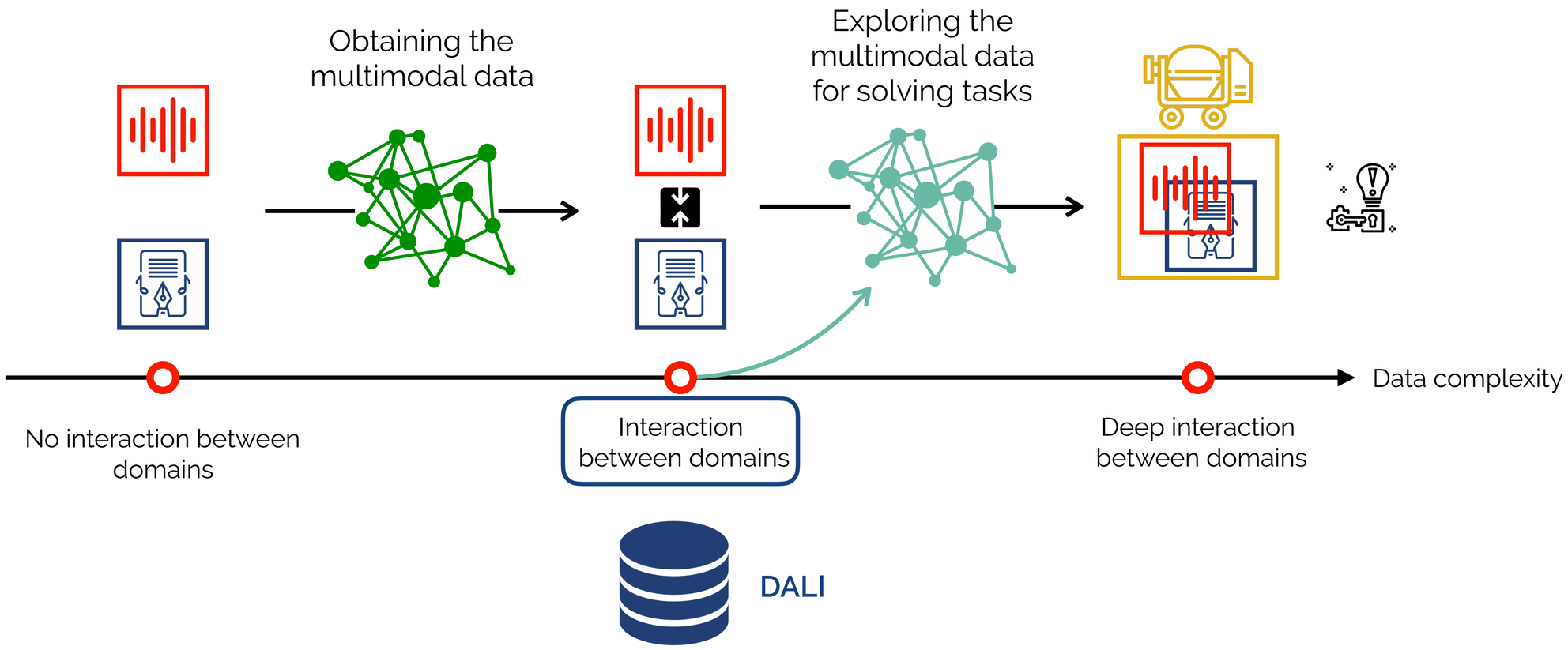
Activation generator

Plan

1. Introduction
2. Dataset of Aligned Lyric Information - DALI
   - 2.1. Motivation
   - 2.2. Creation
   - 2.3. Training with noisy data
3. Multimodal tasks
   - 3.1. Structures analysis
   - 3.2. Source separations
     - 3.2.1.1. Multitasks
     - 3.2.1.2. Vocals
4. Conclusions and future work



Conclusions



Conclusions

1

1. How can we obtain large amounts of labeled data with lyrics aligned in time with the audio?

Gabriel Meseguer-Brocal, Alice Cohen-Hadria and Geoffroy Peeters. (2018) **DALI: a large Dataset of synchronized Audio, Lyrics and notes, automatically created using teacher-student machine learning paradigm**. (ISMIR).

Gabriel Meseguer-Brocal, Alice Cohen-Hadria and Geoffroy Peeters (2020) **Creating DALI, a large dataset of synchronized audio, lyrics, and notes**. (TISMIR).

2. How can we automatically identify errors in these labels?

Gabriel Meseguer-Brocal, Rachel Bittner, Simon Duran, Brian Brost (2020) **Self-Supervised Data Cleansing for Vocal Note Event Annotations.** (ISMIR) Under review

2

3. How can we exploit the relationships between lyrics and audio to improve the performance for lyrics segmentation?

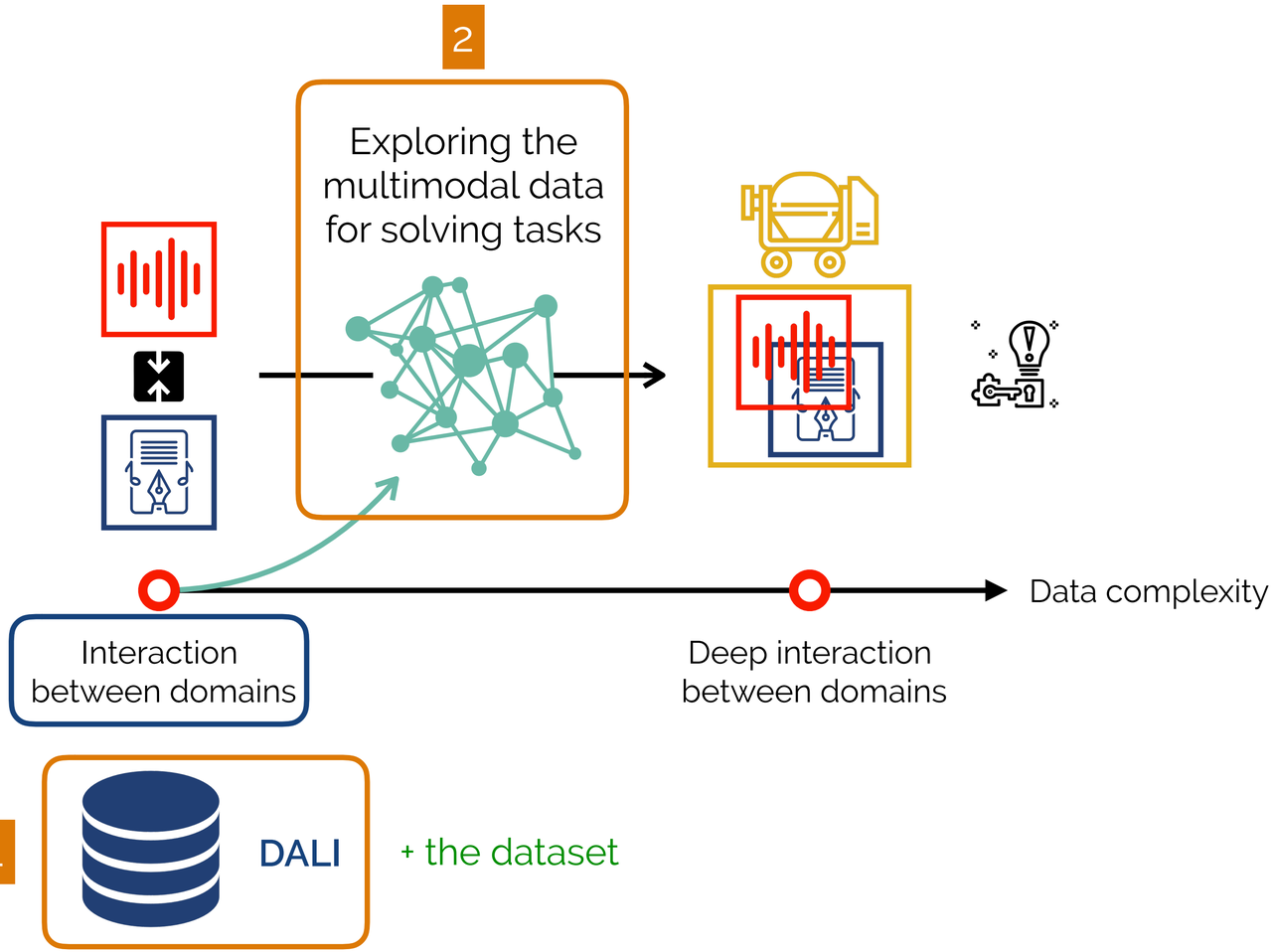
Michael Fell E, Yaroslav Nechaev, **Gabriel Meseguer-Brocal**, Elena Cabrio, Fabien Gandon and Geoffroy Peeters (2020). **Lyrics Segmentation via Bimodal Text-audio Representation**. Natural Language Engineering. Under review

4. How can we control data-driven models by context information?

Gabriel Meseguer-Brocal and Geoffroy Peeters (2019). **Conditioned-U-Net: Introducing a Control Mechanism in the U-Net for Multiple Source Separations** (ISMIR)

5. Can we then use the prior knowledge about the audio signal defined by the lyrics to improve the isolation of the singing voice from the mixture?

Gabriel Meseguer-Brocal and Geoffroy Peeters (2020). **Content based singing voice source separation via strong conditioning using aligned phonemes** (ISMIR) Under review



Future work

1

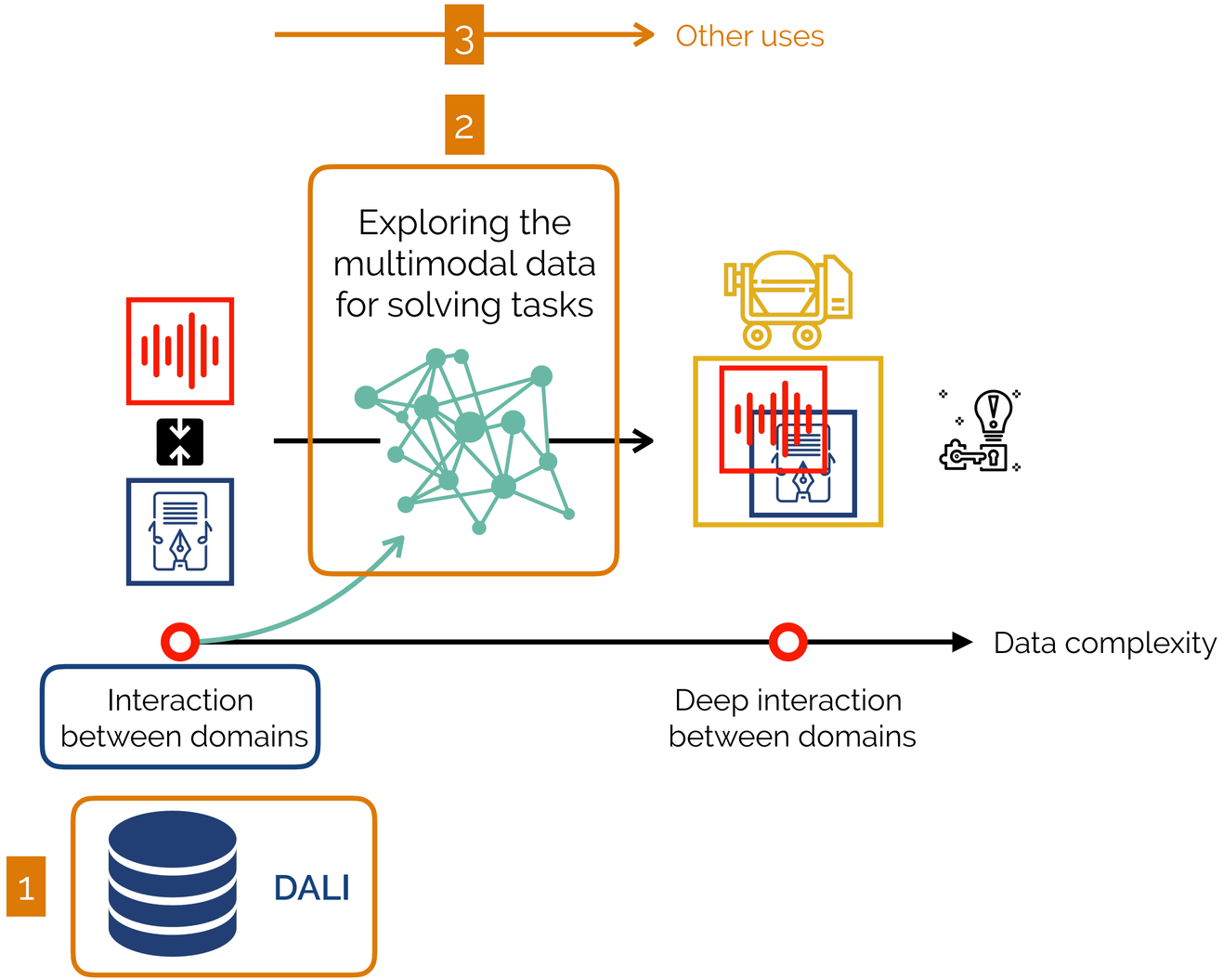
I. To solve the errors using as guide the error function.
II. To add more sources to the multitracks.
III. To perform the alignment per phonemes.

2

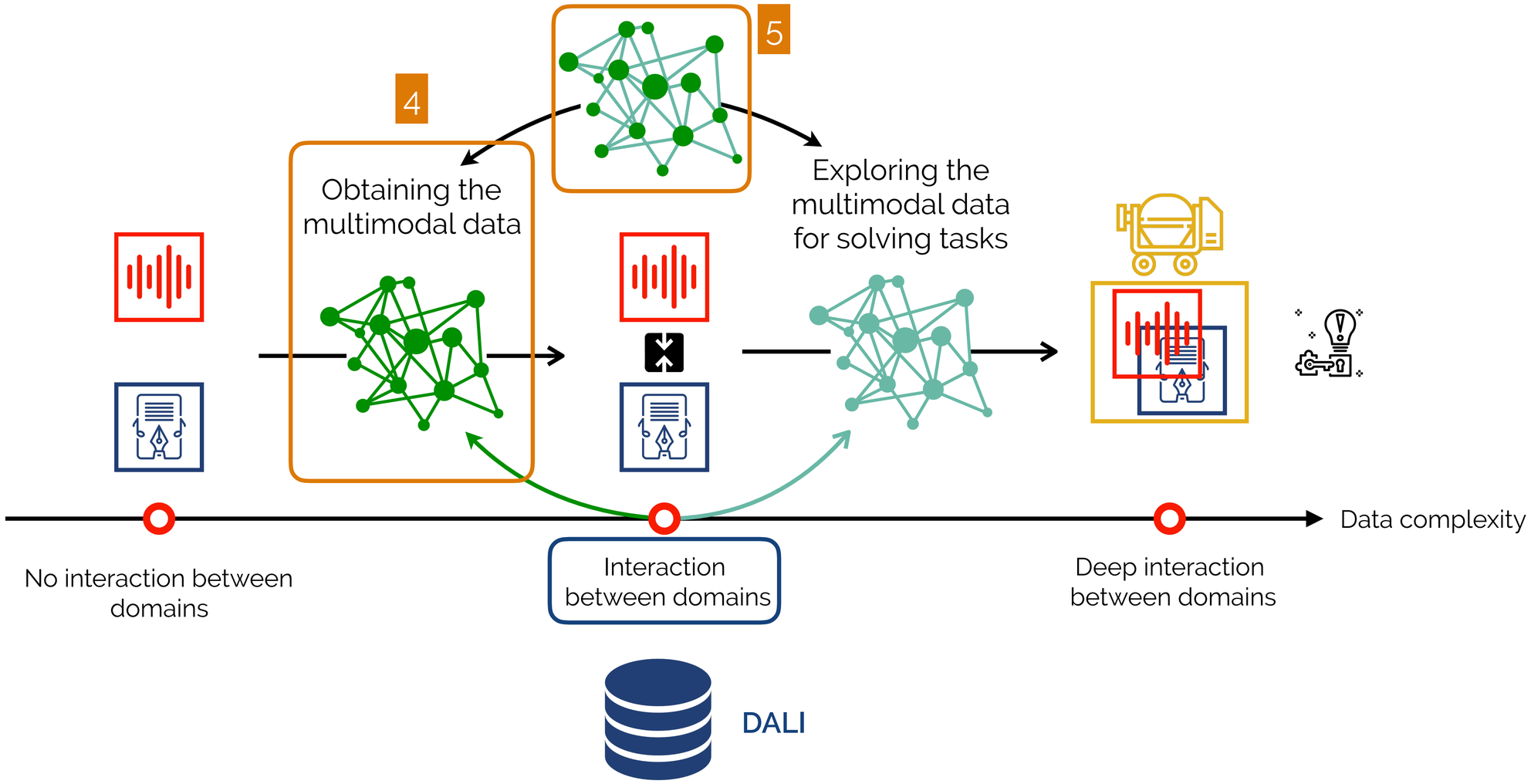
I. To use the text information to help audio-defined segmentation.
II. To test the C-U-Net with more instruments.
III. To explore the conditioning for other architectures.
IV. To use the pitch-notes informations.
V. To introduce attention to be robust to errors in the activation.

3

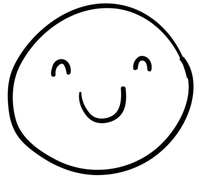
I. Mood estimation.
II. Cover detection.
III. Hierarchy connections between the different granularity levels.

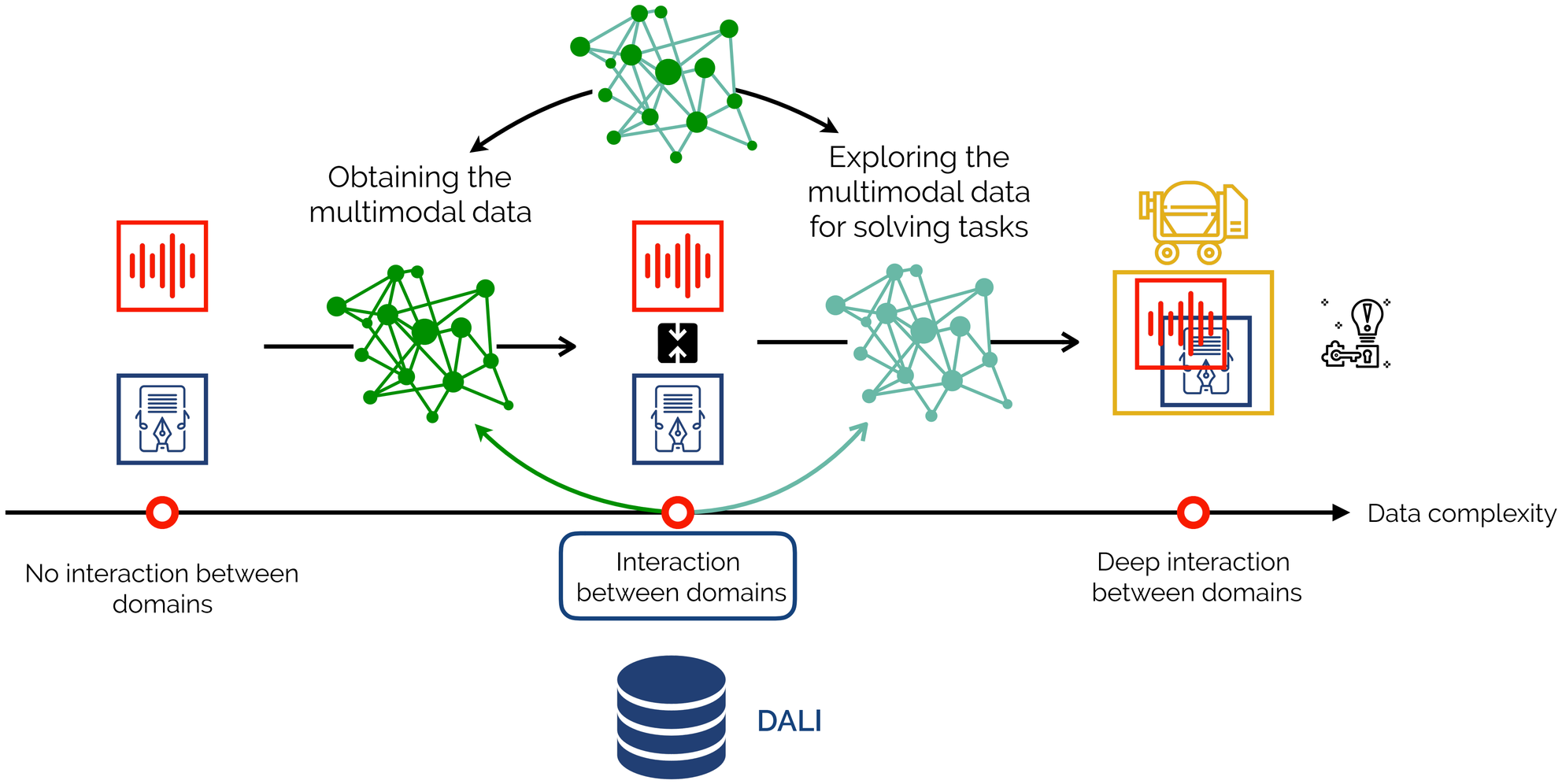


Future work


5
4
Obtaining the multimodal data
Exploring the multimodal data for solving tasks
Data complexity
No interaction between domains
Interaction between domains
Deep interaction between domains
DALI

Thank you!




Obtaining the multimodal data
Exploring the multimodal data for solving tasks
No interaction between domains
Interaction between domains
Deep interaction between domains
Data complexity
DALI