Content based singing voice source separation via strong conditioning using aligned phonemes

Gabriel Meseguer-Brocal
Ircam Lab, CNRS, Sorbonne Université Paris, France
gabriel.meseguerbrocal@ircam.fr

Geoffroy Peeters
LTCI, Télécom Paris, Institut Polytechnique de Paris, Paris, France
geoffroy.peeters@telecom-paris.fr

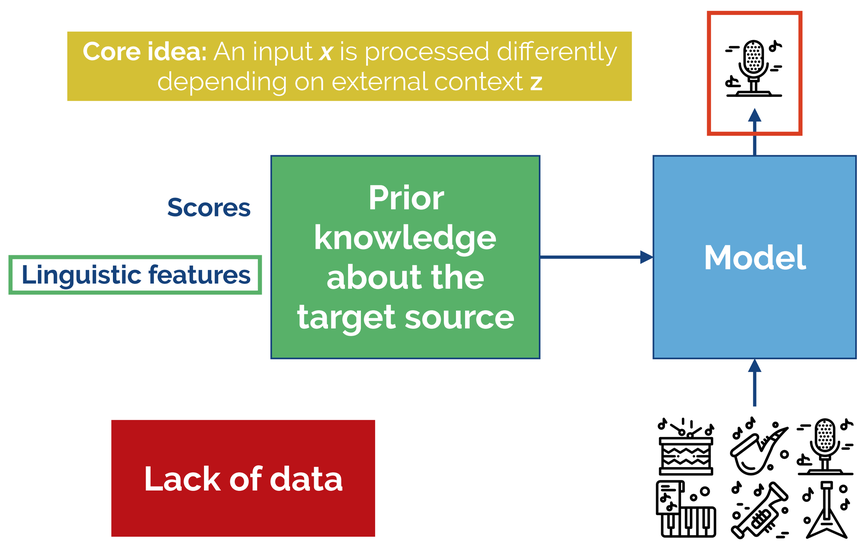
Informed source separation

Source separation aims to isolate instruments from a mix of instruments.

In ***informed source separation***, we use some external prior knowledge about the target source to improve the separation.

Traditionally researchers used scores. However nowadays many have focused on the vocals and used external information that describe the complexity of the vocals signal such as linguistic features.

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



Vocals:

- Central element in popular music.
- Complex signal with a wide range of characteristics depending on the linguistics.

Why?

Add additional flexibility to the model aiming to improve the separation.

Goal - Twofold:

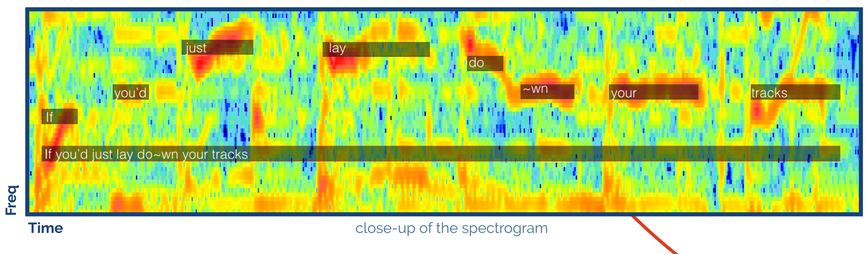
1. to introduce a new multimodal multitrack dataset with lyrics aligned in time.
2. to improve singing voice separation using the prior knowledge defined by the phonetic characteristics.

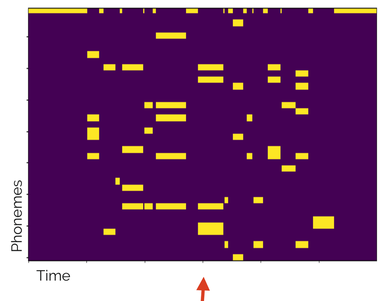
We use the phoneme activation as side information and show that it helps in the separation.

DALI Multitracks

The mixture is divided into vocals and accompaniment and each multitrack has its lyrics aligned in time at four levels of granularity

Sources	*Songs*	*Artists*	*Geners*	*Decades*	*Laguange*
Vocals + accompaniment	513	247	32	7	English






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

Multitrack version of the original DALI:

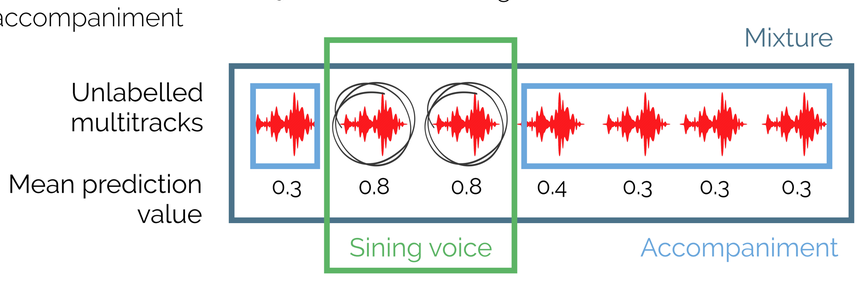
Gabriel Meseguer-Brocal, Alice Cohen-Hadria and Geoffroy Peeters. (2018) **DALI: a large Dataset of synchronized Audio, Lyrics and notes, automatically created using teacher-student machine learning paradigm**. (ISMIR).

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

Method for creating the vocals, accompaniment and mixture stems from the raw stems.

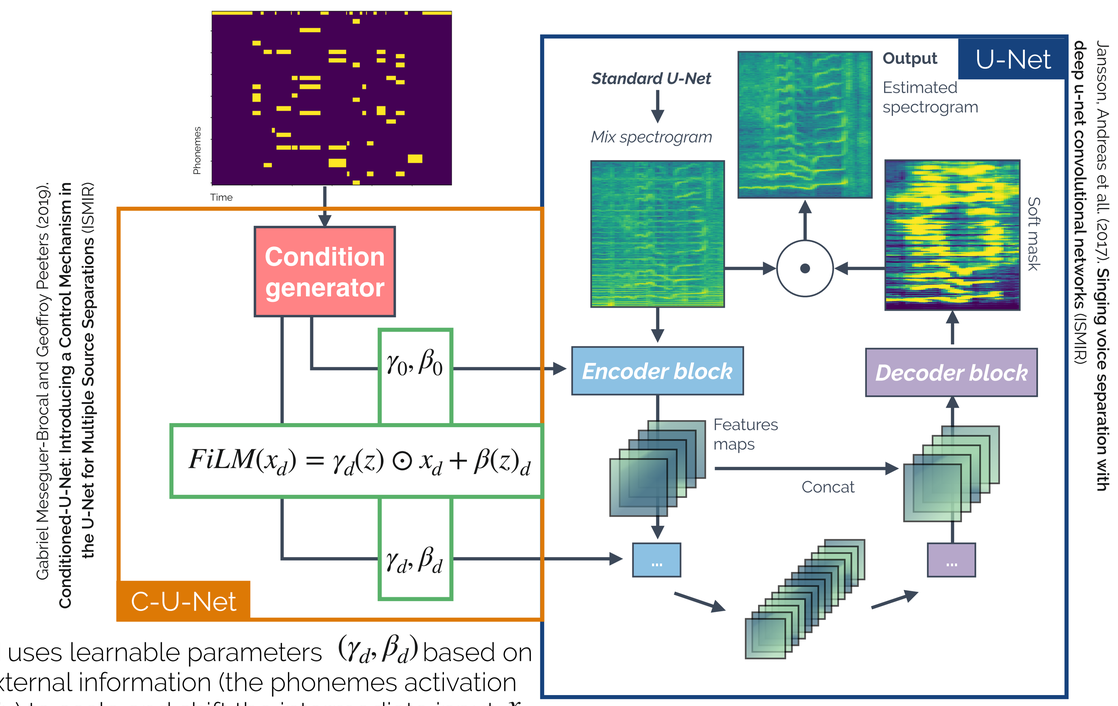
We compute a singing voice probability vector overtime, using a pretrained Singing Voice Detection (SVD) model.
We obtain then a global mean prediction value per tracks. τ .

Assuming that there is at least one track with vocals, we create the vocals source by merging all the tracks with $\epsilon_\tau >= \max_\tau(\epsilon_\tau) \cdot \nu$ where ν is a tolerance value set to 0.98. All the remaining tracks are fused to define the accompaniment



Conditioned source separation

Weak conditioning Same FiLM operation to the whole input patch.



FiLM uses learnable parameters (γ_d, β_d) based on an external information (the phonemes activation matrix) to scale and shift the intermediate input x_d

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

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

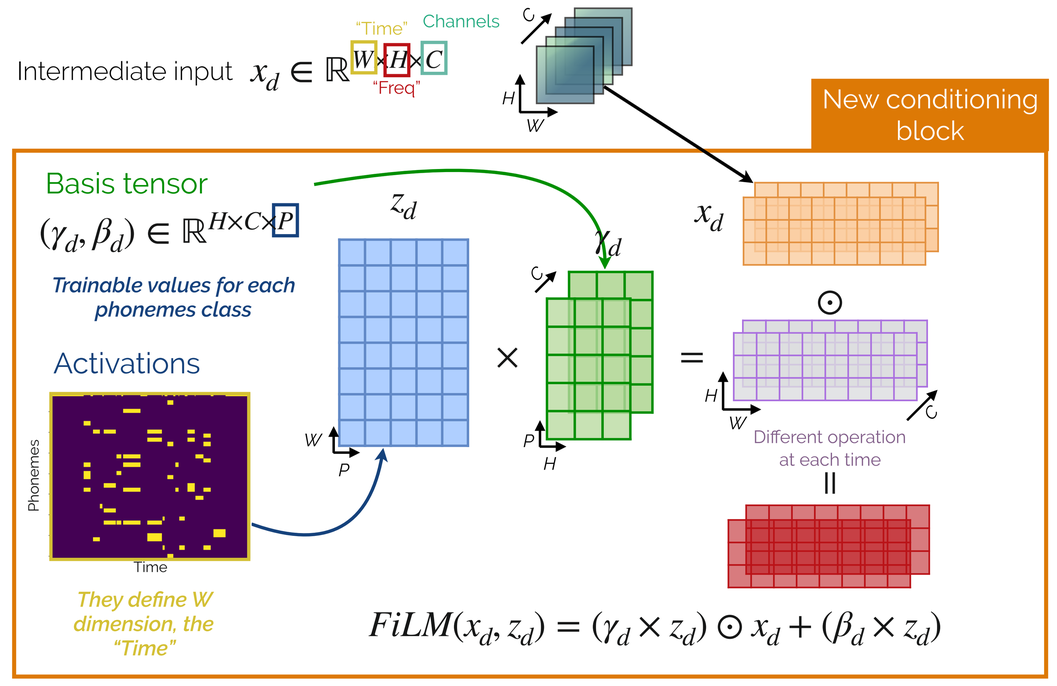
We are loosing the time information!!

We are embedding the activation to apply the a single FiLM operation to the whole input patch.

Vs

Strong conditioning Different FiLM operations at different time frames.

Intermediate input $x_d \in \mathbb{R}^{W \times H \times C}$ ("Time", "Freq", Channels)

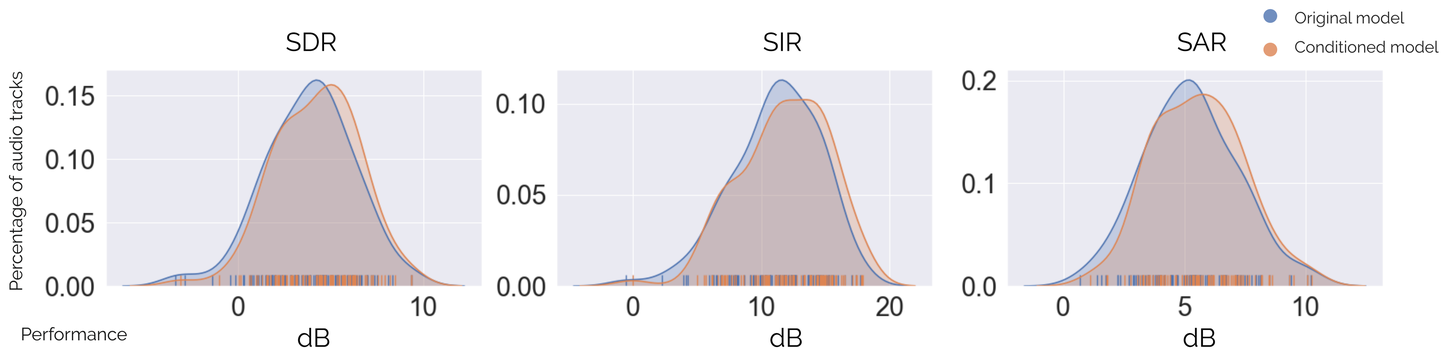


Note: the conditioned net is still a U-Net, this replaces the control mechanism

Results

Model	SDR	SIR	SAR	PES	EPS
U-Net	4.05	11.40	5.32	-42.44	-64.84
W_{si}	**4.24**	**11.78**	5.38	**-49.44**	-65.47
W_{co}	**4.24**	**12.72**	5.15	**-59.53**	-63.46
S_a	4.04	**12.14**	5.13	**-59.68**	-61.73
S_{a*}	**4.27**	**12.42**	5.26	**-54.16**	-64.56
S_c	**4.36**	**12.47**	5.34	**-57.11**	-65.48
S_{c*}	**4.32**	**12.86**	5.15	**-54.27**	**-66.35**
S_f	4.10	11.40	5.24	47.75	-62.76
S_{f*}	4.21	**13.13**	5.05	**-48.75**	**-72.40**
S_s	**4.45**	**11.87**	**5.52**	**-51.76**	-63.44
S_{s*}	**4.26**	**12.80**	5.25	**-57.37**	-65.62

Median performance in dB of the different mod-els on the DALI test set. In bold are the results that significantly improve over the U-Net ($p < 0.001$) and inside the circles the best results for each metric



Distribution of scores for the the standard U-Net (Blue) and S_s (Orange) with $(\gamma_d, \beta_d) \in \mathbb{R}^P$

This research has received funding from the French National Research Agency under the contract ANR-16-CE23-0017-01 (WASABI project)

Code at: **https://github.com/gabolsgabs/vunet**
Database at: **https://zenodo.org/deposit/3970189**