Jointly Detecting and Separating Singing Voice: A Multi-Task Approach

Daniel Stoller[1], Sebastian Ewert[2*], Simon Dixon[1]

[1]Centre for Digital Music
Queen Mary University of London

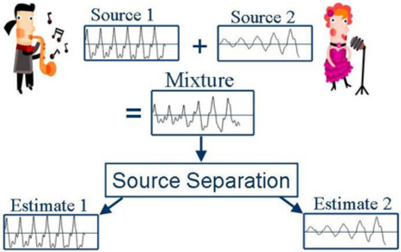
[2]Spotify
London

LVA ICA
05.07.2018

*Work was conducted at Queen Mary University of London

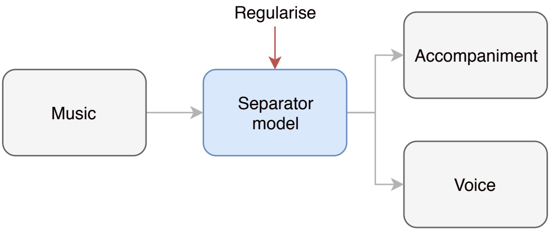
Vocal separation
Introduction

- Main task: Separate vocals from music pieces
- Applications: Karaoke generation, singer identification, voice analysis...



Vocal separation
Challenges

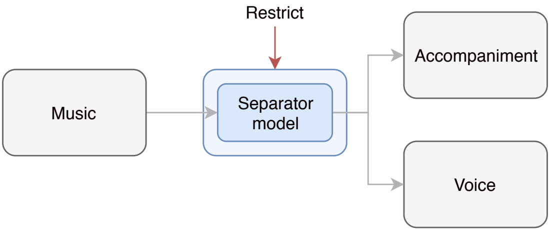
- Difficult task, small multi-track datasets ⇒ Overfitting
- Give model more knowledge:
  - Regularization (e.g. weight decay)



Vocal separation

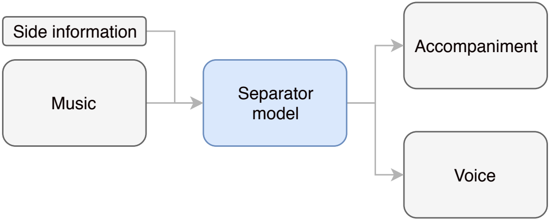
Challenges

- Difficult task, small multi-track datasets $\Rightarrow$ Overfitting
- Give model more knowledge:
  - Knowledge-driven (e.g. KAM [4])



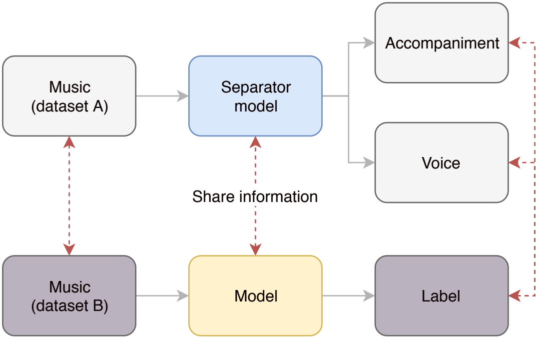
Vocal separation
Challenges

- Difficult task, small multi-track datasets $\Rightarrow$ Overfitting
- Give model more knowledge:
  - Informed source separation [2]



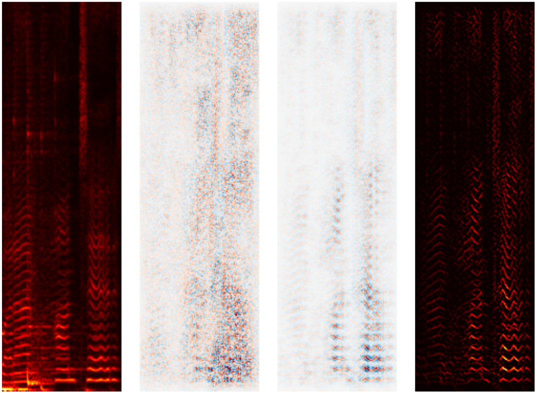
Vocal separation
Challenges

- Difficult task, small multi-track datasets ⇒ Overfitting
- Give model more knowledge:
  - **Integrate information from related tasks/datasets**



Goal

- Which other tasks could help?
- Vocal activity detection is promising:
  - Knowing vocal activity improves vocal separation [1]
  - Vocal detection networks learn a form of separation: [5]



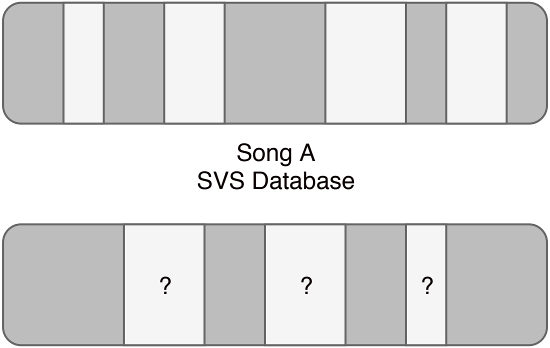
(a) (b) (c) (d)

Initial approach

Using additional non-vocal sections

- U-Net adaptation [3] as separator, MSE loss
- **Sample instrumental sections also from SVD databases**

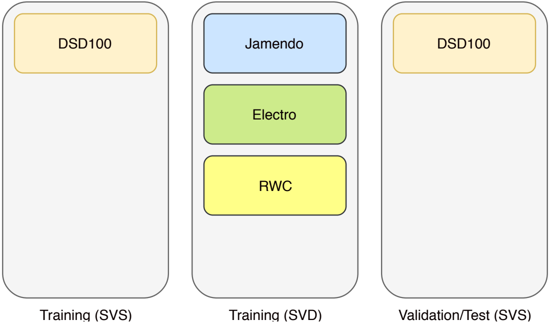
⇒ Diversifies instrumental training data



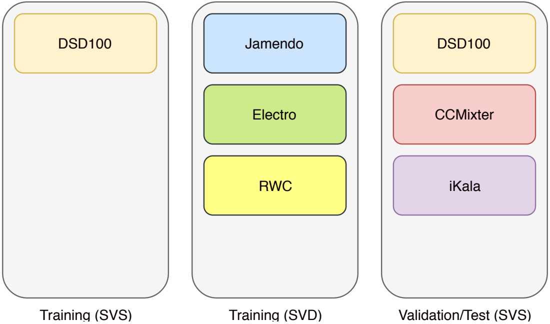
Song A
SVS Database

Song B
SVD Database

Initial approach
Results

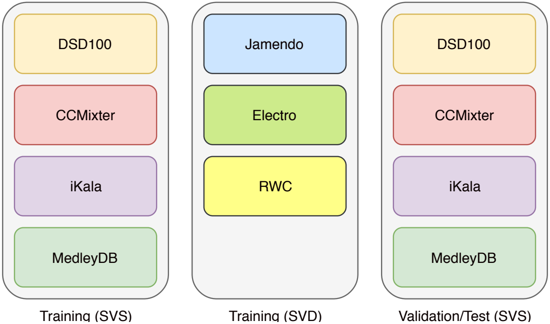


Performance **decrease**



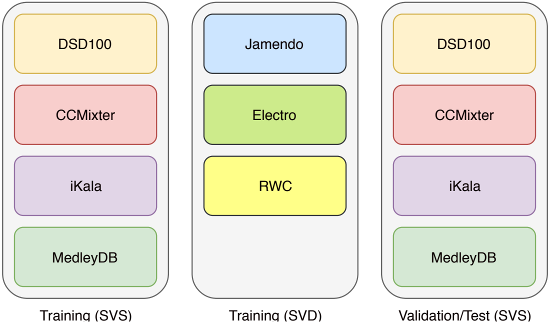
Performance **increase**

Initial approach
Results



Performance **decrease**

Initial approach
Results

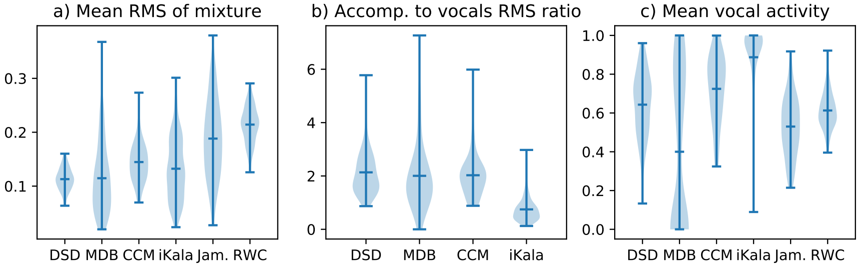


Performance **decrease**

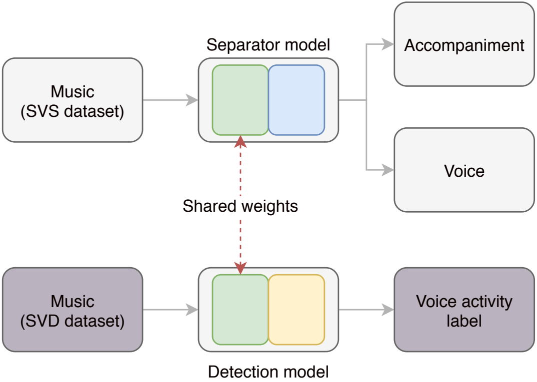
Dataset bias?

Dataset bias

Analysis



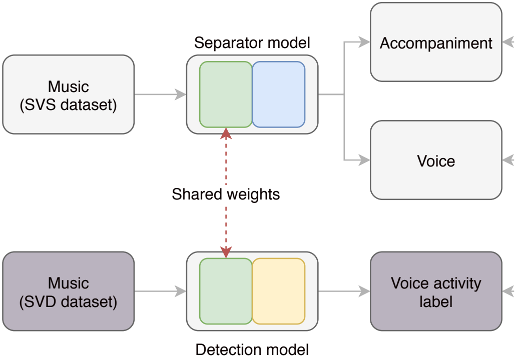
Key idea: Predict both audio and label



Multi-task approach

Introduction and motivation

Key idea: Predict both audio and label



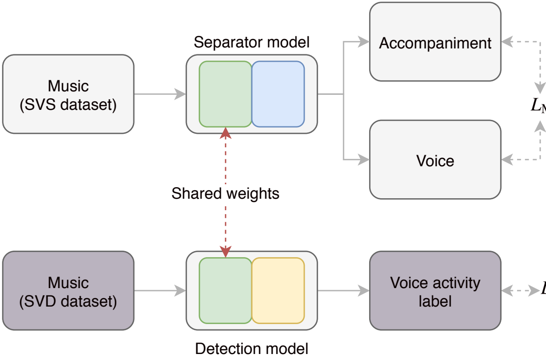
$$L_{\mathrm{MSE}} = \mathbb{E}_{(\mathbf{m},\mathbf{s})\sim p^1} \frac{1}{N} ||\mathbf{s} - f_\phi(\mathbf{m})||^2$$

$$L_{\mathrm{CE}} = \mathbb{E}_{(\mathbf{m},\mathbf{o})\sim p^2} \frac{1}{T} \sum_{t=1}^{T} \log p_\phi^t(o_t|\mathbf{m})$$

$$L_{\mathrm{MTL}} = \alpha L_{\mathrm{MSE}} + (1-\alpha) L_{\mathrm{CE}}$$

Multi-task approach
Introduction and motivation

Key idea: Predict both audio and label



$$L_{\text{MSE}} = \mathbb{E}_{(\mathbf{m},\mathbf{s})\sim p^1} \frac{1}{N} ||\mathbf{s} - f_\phi(\mathbf{m})||^2$$

$$L_{\text{CE}} = \mathbb{E}_{(\mathbf{m},\mathbf{o})\sim p^2} \frac{1}{T} \sum_{t=1}^{T} \log p_\phi^t(o_t|\mathbf{m})$$

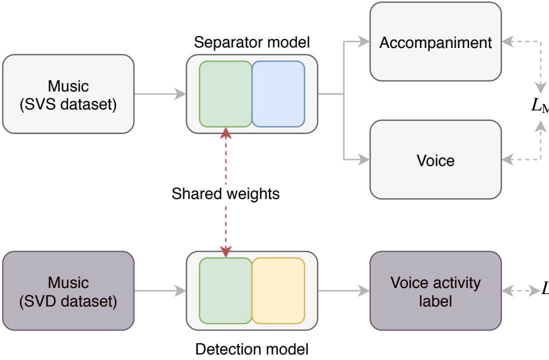
$$L_{\text{MTL}} = \alpha L_{\text{MSE}} + (1-\alpha) L_{\text{CE}}$$

Robust to dataset bias and label accuracy

Multi-task approach

Introduction and motivation

Key idea: Predict both audio and label



$$L_{\mathrm{MSE}} = \mathbb{E}_{(\mathbf{m},\mathbf{s})\sim p^1} \frac{1}{N} ||\mathbf{s} - f_\phi(\mathbf{m})||^2$$

$$L_{\mathrm{CE}} = \mathbb{E}_{(\mathbf{m},\mathbf{o})\sim p^2} \frac{1}{T} \sum_{t=1}^{T} \log p_\phi^t(o_t|\mathbf{m})$$

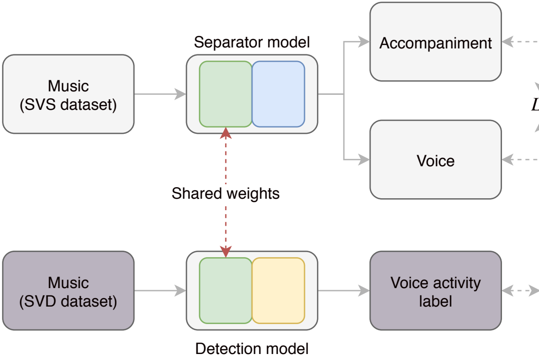
$$L_{\mathrm{MTL}} = \alpha L_{\mathrm{MSE}} + (1-\alpha) L_{\mathrm{CE}}$$

Can train with vocal sections from SVD data

Multi-task approach

Introduction and motivation

Key idea: Predict both audio and label



$$L_{\text{MSE}} = \mathbb{E}_{(\mathbf{m},\mathbf{s})\sim p^1} \frac{1}{N} ||\mathbf{s} - f_\phi(\mathbf{m})||^2$$

$$L_{\text{CE}} = \mathbb{E}_{(\mathbf{m},\mathbf{o})\sim p^2} \frac{1}{T} \sum_{t=1}^{T} \log p_\phi^t(o_t|\mathbf{m})$$

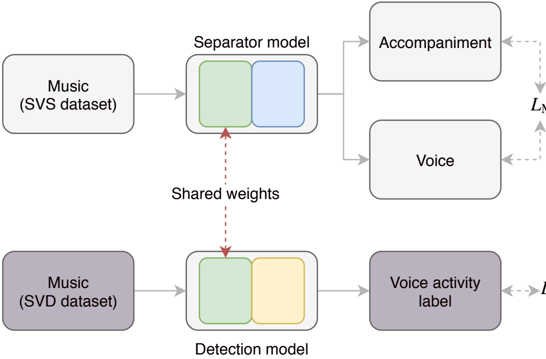
$$L_{\text{MTL}} = \alpha L_{\text{MSE}} + (1-\alpha) L_{\text{CE}}$$

Needs only mixture at test time

Multi-task approach

Introduction and motivation

Key idea: Predict both audio and label



$$L_{\mathrm{MSE}} = \mathbb{E}_{(\mathbf{m},\mathbf{s})\sim p^1} \frac{1}{N} ||\mathbf{s} - f_\phi(\mathbf{m})||^2$$

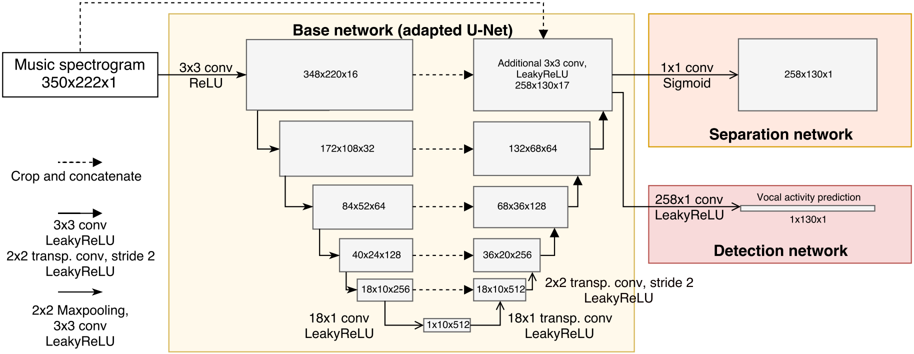
$$L_{\mathrm{CE}} = \mathbb{E}_{(\mathbf{m},\mathbf{o})\sim p^2} \frac{1}{T} \sum_{t=1}^{T} \log p^t_\phi(o_t|\mathbf{m})$$

$$L_{\mathrm{MTL}} = \alpha L_{\mathrm{MSE}} + (1-\alpha) L_{\mathrm{CE}}$$

Solves two tasks at once

Experimental setup

Model architecture and dataset



- DSD100 as SVS, Jamendo as SVD training data

Experimental setup

Evaluation metrics: AU-ROC, MSE, SDR

- AU-ROC for SVD
- MSE and SDR/SIR/SAR for separation
  - SDR gives log(0) for non-vocal sections ($\approx 10\%$)

  $\Rightarrow$ Also measure RMS of vocal estimates for non-vocal sections

Results

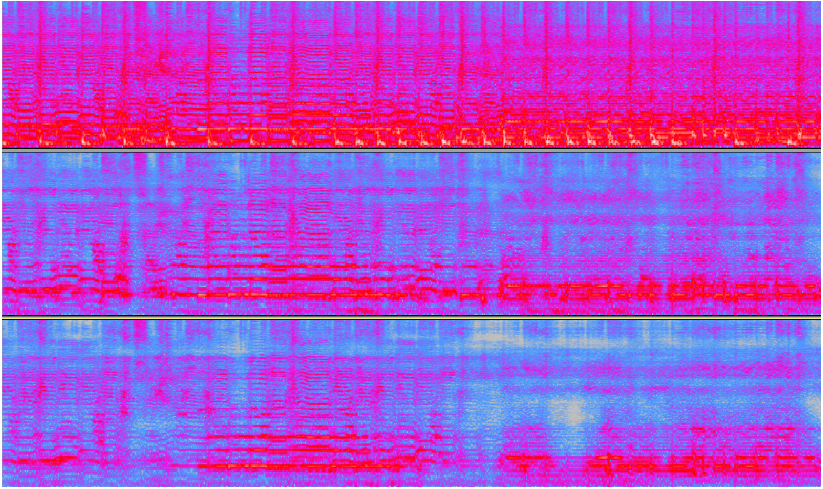
Single-task vs. multi-task model

		Metric								
					Vocals			Accompaniment		
		AU-ROC	MSE	Non-voc. RMS	SDR	SIR	SAR	SDR	SIR	SAR
Model	SVD	0.9239	-	-	-	-	-	-	-	-
	SVS	-	0.01865	0.0194	2.83	5.27	**6.88**	6.71	**14.75**	13.25
	Ours	**0.9250**	**0.01755**	**0.0155**	2.86	**5.56**	6.23	6.69	13.24	**14.11**

Table: Comparing SVS and SVD baseline with our approach

Results

Qualitative comparison



Summary

- Current SotA methods only use multi-track data
- Our approach also uses SVD databases
- Improved separation and detection performance
- Future work: Larger datasets, more related tasks

T.-S. Chan, T.-C. Yeh, Z.-C. Fan, H.-W. Chen, L. Su, Y.-H. Yang, and R. Jang.
Vocal activity informed singing voice separation with the ikala dataset.
In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 718–722. IEEE, 2015.

S. Ewert, B. Pardo, M. Müller, and M. D. Plumbley.
Score-informed source separation for musical audio recordings: An overview.
IEEE Signal Processing Magazine, 31(3):116–124, 2014.

A. Jansson, E. J. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde.
Singing voice separation with deep U-Net convolutional networks.
In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 323–332, 2017.

A. Liutkus, D. Fitzgerald, and Z. Rafii.
Scalable audio separation with light kernel additive modelling.

In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 76–80. IEEE, 2015.

J. Schlüter.
Learning to pinpoint singing voice from weakly labeled examples.
In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 44–50, 2016.