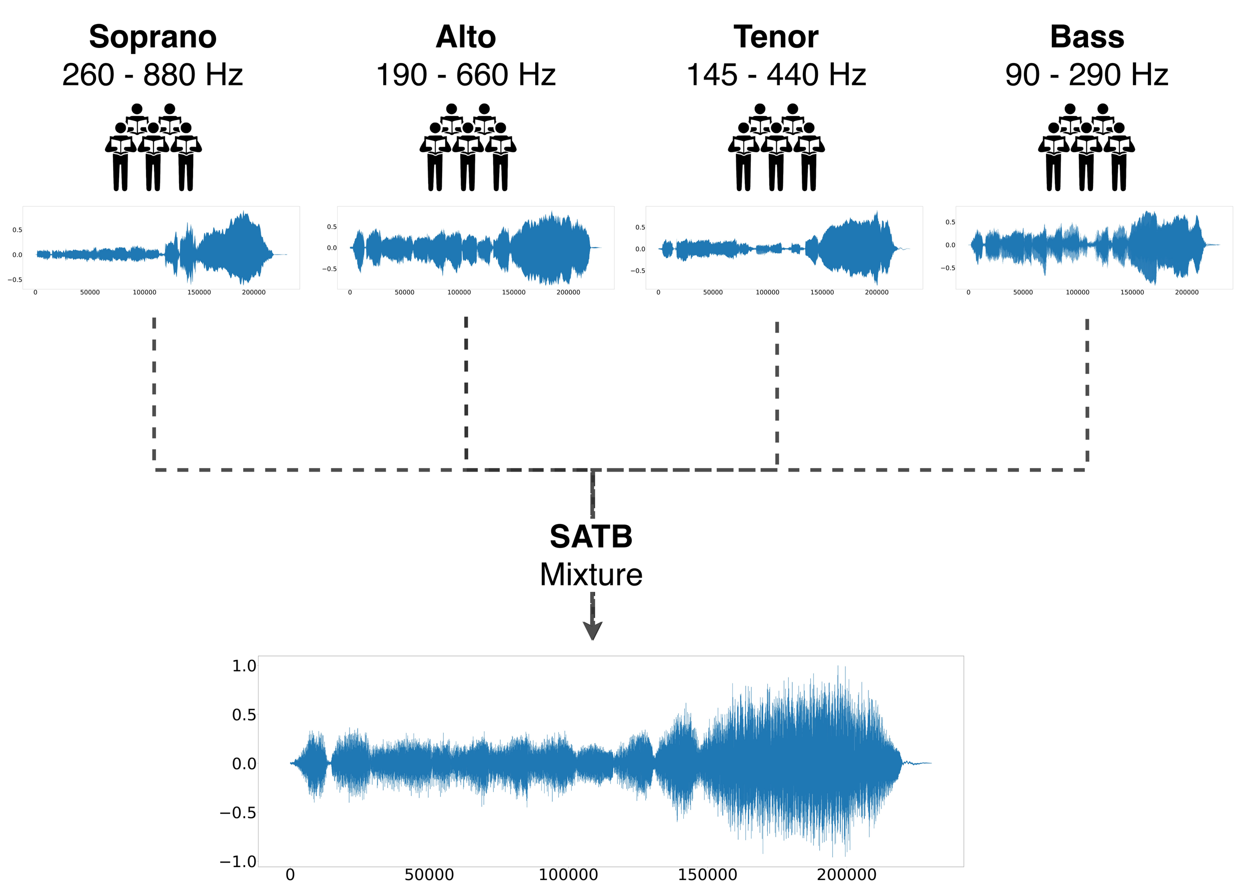
Deep Learning Based Source Separation Applied to Choir Ensembles

Darius Pétermann[1], Pritish Chandna[1], Helena Cuesta[1], Jordi Bonada[1], Emilia Gómez[1, 2]

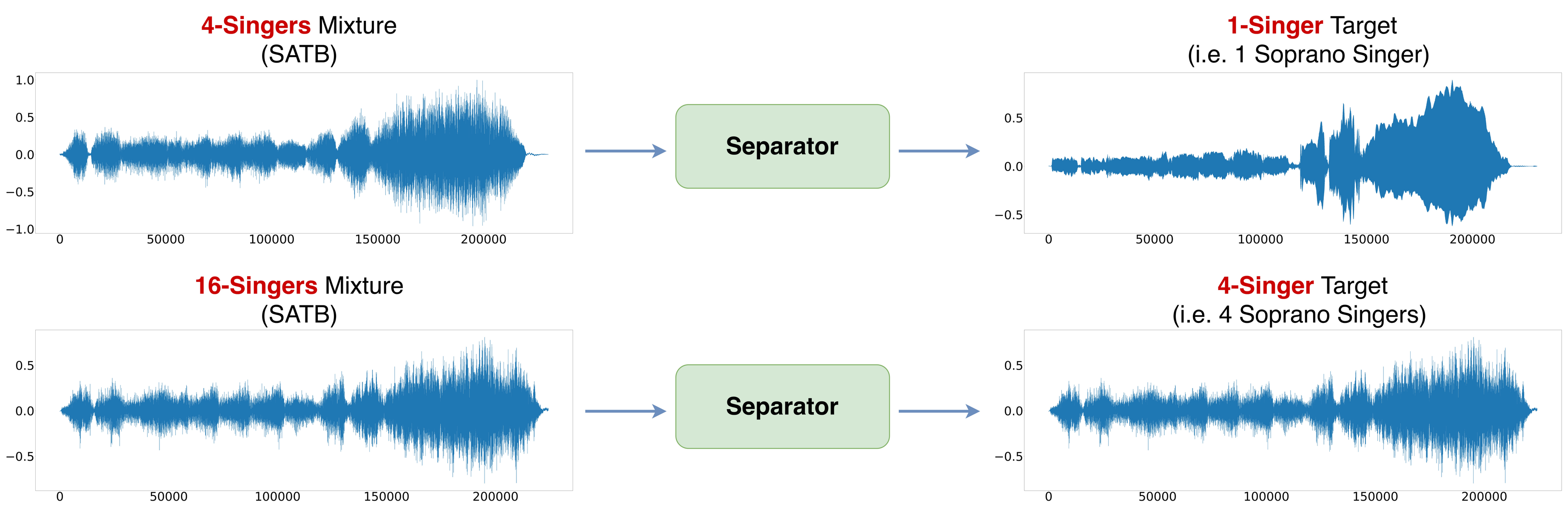
[1] *Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain*
[2] *Joint Research Centre, European Commission, Seville, Spain*

Background



- **SATB** is a common type of choral setting.
- **High correlation** between the signals to be separated makes the separation process challenging.
- Each singing group performs within its own **frequency range**.

Task and Use-Cases



- The task consists in **isolating** each of the four SATB singing groups from a given choir mixture.
- The task is divided into two different use-cases:
  - Use-case 1: Involves **4-singers** mixtures for **1 singer** exactly per singing part.
  - Use-case 2: Involves **16-singers** mixtures for **4 singers** exactly per singing part.

State-of-the-Art & Adaptations

Models	Domain-Agnostic (D-A)	Domain-Specific (D-S)	Domain
Wave-U-Net [1]	✓		Waveform
U-Net [2]	✓		Spectrogram
Open-Unmix [3]	✓		Spectrogram
Conditioned-U-Net D-A [4]	✓		Informed Spec.
Conditioned-U-Net D-S Local		✓	Informed Spec.
Conditioned-U-Net D-S Global		✓	Informed Spec.

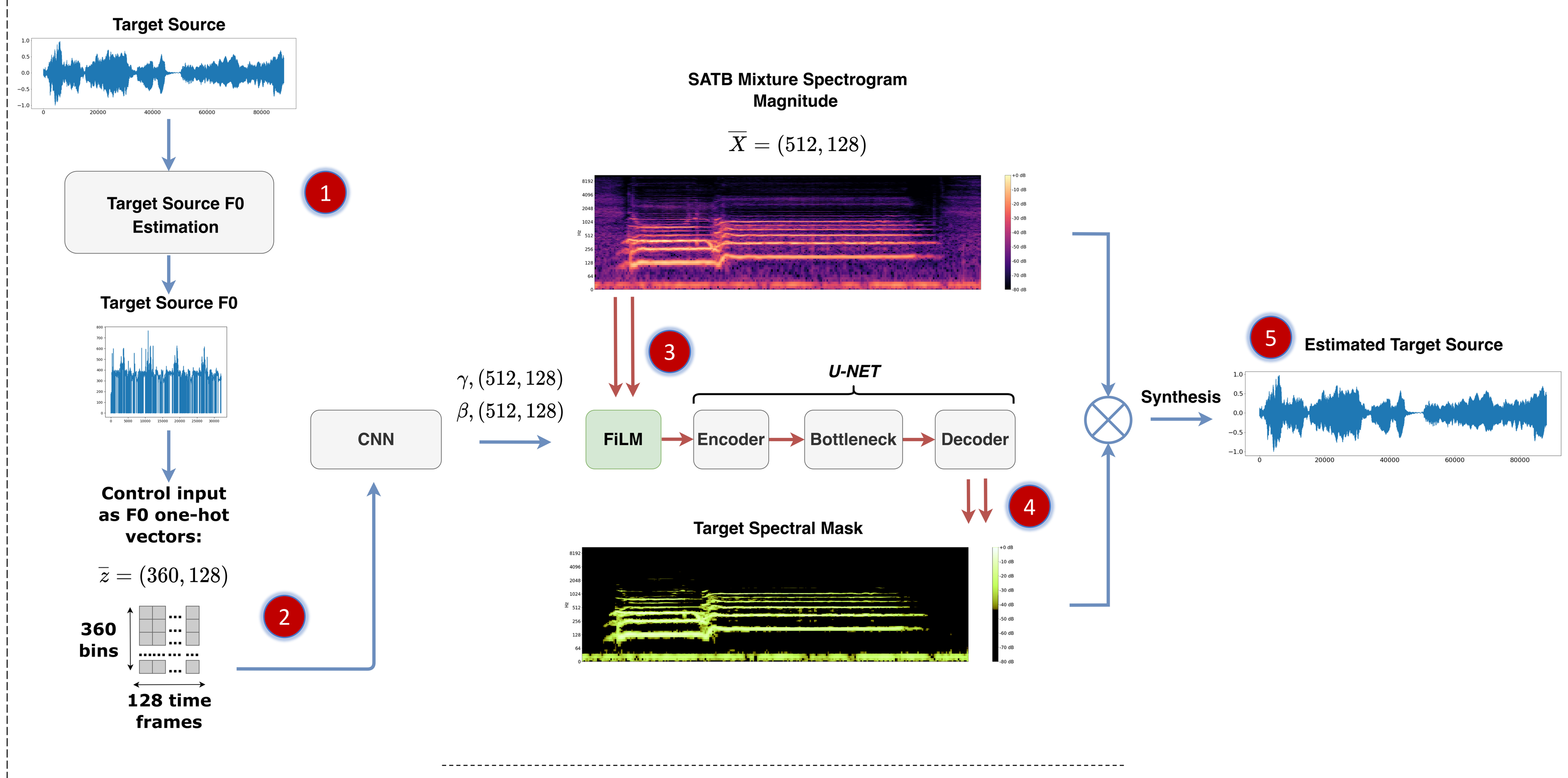
- Recent deep learning architectures used for musical source separation are evaluated, specifically on our task. These models are referred to as "**domain-agnostic**", or "**D-A**".
- Two direct adaptations of the *Conditioned-U-Net* are then proposed (denoted in red). These adaptations consider information conveyed by the sources (i.e. **F0 track**) to improve the separation process; they are described as "**domain-specific**", or "**D-S**".

Dataset & Train-Test Split

- The ***Choral Singing Dataset***, containing 3 songs for 16 stems per songs (4 singers per singing group), as well as a **proprietary dataset** consisting of 25 songs for 4 stems per song (1 singer per singing group), were used for training and testing.
- Due to the limiting nature of the datasets, **training** was performed using portions of both datasets on a **4-singers mixture-basis**, for both use-cases.
- **Testing** was performed as follows for the two use-cases:
  - Use-Case 1: One singer per singing group was set aside for evaluation.
  - Use-Case 2: The entirety of the *Choral Singing Dataset* was used for evaluation.

Universitat Pompeu Fabra Barcelona — MTG Music Technology Group

Domain-Specific Conditioned-U-Net



1. Target source's **F0 track** is obtained through an F0 estimation algorithm [5].
2. The F0 track is then converted into a **2-D one-hot matrix** and input into a CNN.
3. The input spectrogram is transformed by **the set of scalars** output by the CNN.
4. The U-Net output the target's **spectral mask**.
5. The predicted target source is **synthesized** using the resulting magnitude spectrogram and the phase from the input mixture.

Objective Evaluation: BSS Eval

Model	Test Use-Case 1 - SDR (dB) Soprano	Alto	Tenor	Bass	Avg.
Wave-U-Net	2.03±2.2	4.59±2.7	0.92±2.9	2.72±2.5	2.56±2.3
U-Net	3.78±2.1	5.15±3.7	2.29±2.7	3.22±1.5	3.61±2.5
C-U-Net D-A	3.57±2.0	2.05±2.1	-1.25±2.6	1.96±2.2	1.58±2.2
Open-Unmix	5.61±2.1	5.70±2.3	1.60±1.7	3.66±2.2	4.14±2.1
C-U-Net D-S L	3.70±1.3	6.99±1.9	3.82±1.6	3.74±1.7	4.56±1.6
C-U-Net D-S G	**5.76**±1.2	**7.67**±1.5	**5.39**±1.4	**4.07**±1.8	**5.73**±1.5

Model	Test Use-Case 2 - SDR (dB) Soprano	Alto	Tenor	Bass	Avg.
Wave-U-Net	3.30±1.6	4.73±0.8	2.09±2.0	1.24±1.4	2.84±1.5
U-Net	5.14±1.5	6.63±1.0	4.74±1.7	3.12±1.6	4.91±1.4
C-U-Net D-A	4.61±1.8	2.67±2.7	0.52±2.8	1.98±1.6	2.45±2.2
Open-Unmix	**6.67**±2.1	6.49±1.3	2.70±1.6	**3.49**±2.0	4.83±1.7
C-U-Net D-S L	4.34±0.9	**7.06**±1.2	4.77±1.6	3.48±1.5	4.91±1.3
C-U-Net D-S G	5.34±1.2	6.44±1.4	**4.93**±1.5	3.18±1.1	**4.97**±1.3

Extensive BSS Eval results (SDR, SIR, SAR) as well as audio examples can be found at the following: *https://darius522.github.io/satb-source-separation-results/*

Results and Discussion

- Introducing **domain-knowledge** (i.e. F0 track) during training and inference improves the model's separation performance on the four SATB parts for both use-cases.
- The improvement gap between domain-agnostic and domain-specific models is less evident on **the second use-case**. This could be explained by the fact that **the mean of the various pitches** present in a singing group is not necessarily representative of the true underlying pitch of the unison.

References

[1] D. Stoller, S. Ewert, and S. Dixon, (2018) "**Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation**"

[2] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, (2017) "**Singing Voice Separation with Deep U-Net Convolutional Networks**"

[3] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, (2019) "**Open-unmix - a reference implementation for music source separation**"

[4] Meseguer-Brocal, Gabriel & Peeters, Geoffroy. (2019). **Conditioned-U-Net: Introducing a Control Mechanism in the U-Net for Multiple Source Separations**.

[5] H. Cuesta et al., (2020) "**Multiple F0 Estimation in Vocal Ensembles using Convolutional Neural Networks**"

TROMPA