Score-Informed Source Separation of Choral Music

Matan Gover and Philippe Depalle

McGill · Schulich School of Music / École de musique Schulich

Code, dataset, and audio examples:
https://www.matangover.com/choirsep

Task

input: choir recording

output: individual track for each choir section

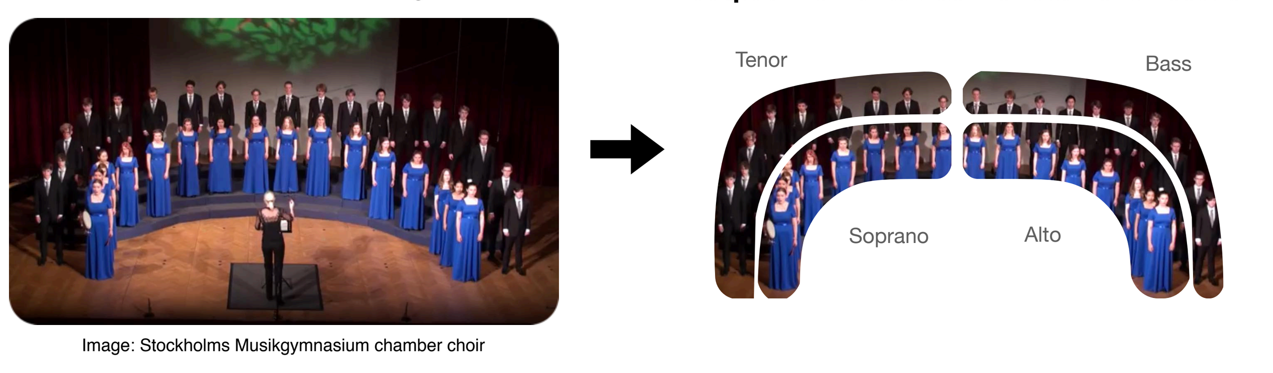


Image: Stockholms Musikgymnasium chamber choir

Challenges

- Separation must undo the "choral blend"
- Choral timbre is complex: each choir section is composed of multiple singers with pitch, timing, and timbre variations
- Lack of publicly available datasets for training
- New task: no baselines for comparison

Synthesized Multi-Track Choir Dataset

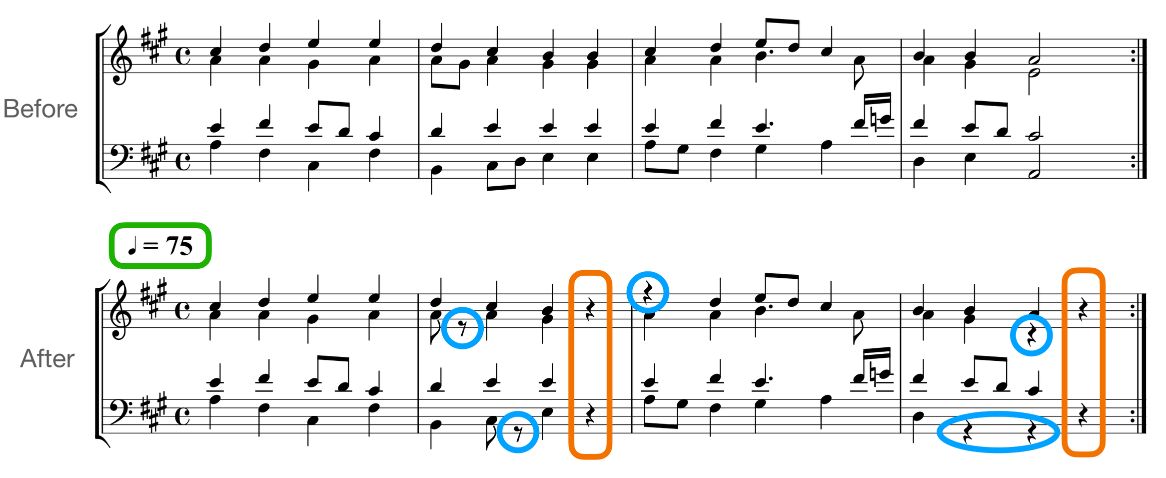
Due to the lack of publicly available datasets for training, we create our own dataset by synthesizing 351 chorale harmonizations by J. S. Bach.

For synthesis we use FluidSynth with a 'Choir Aahs' preset (SoundFont distributed with MuseScore). Unfortunately cannot synthesize lyrics.

Total dataset duration: 3h 48m

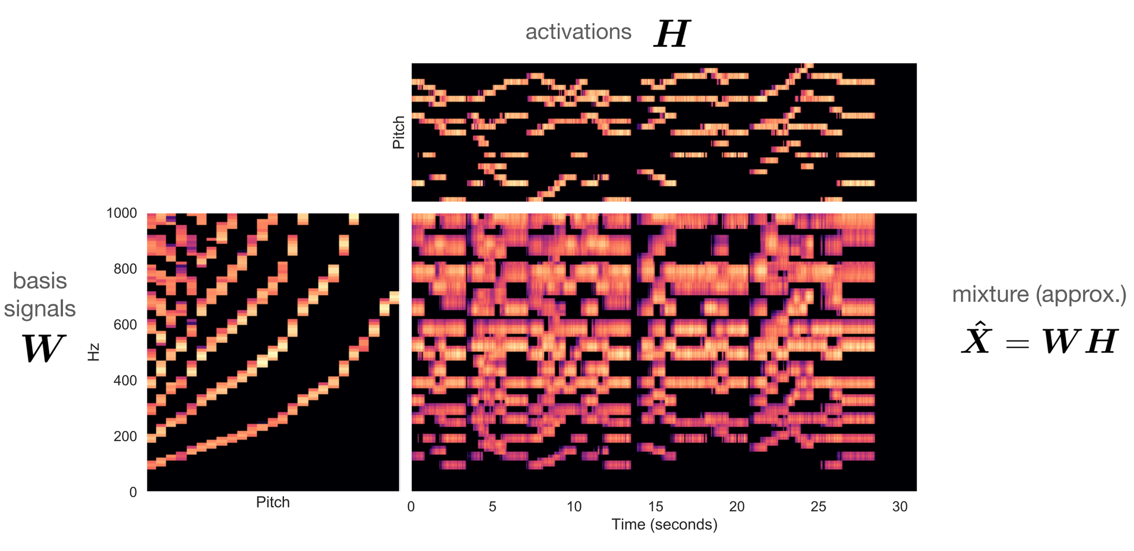
Data Augmentation

simulated breaths, random omitted notes, and tempo variations



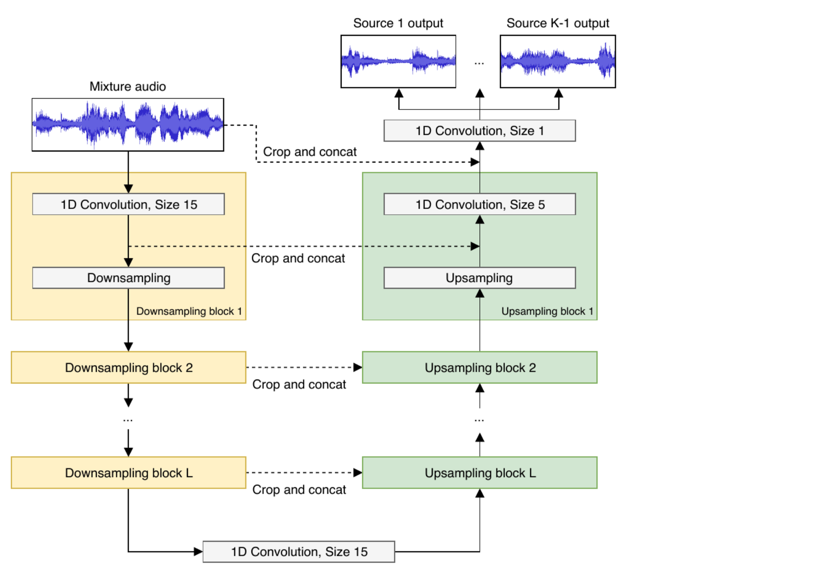
Baseline: Score-Informed NMF

- NMF (non-negative matrix factorization) produces an approximate factorization of a mixture spectrogram as a product of two matrices: basis signals and activations
- Basis signals and activations are constrained using the musical score [1]
- Disadvantage: Factorization is discrete, so cannot properly account for continuous evolution of spectral parameters over time. This leads to artifacts in separation results



Wave-U-Net

- Convolutional neural network, originally for vocals & accompaniment separation [2]
- Operates directly on the time-domain signal
- Encoder-decoder architecture with skip connections:
  - gradually downsamples the input signal to a low-resolution bottleneck
  - upsamples from the bottleneck back to the original signal resolution, using also the output of each downsampling layer

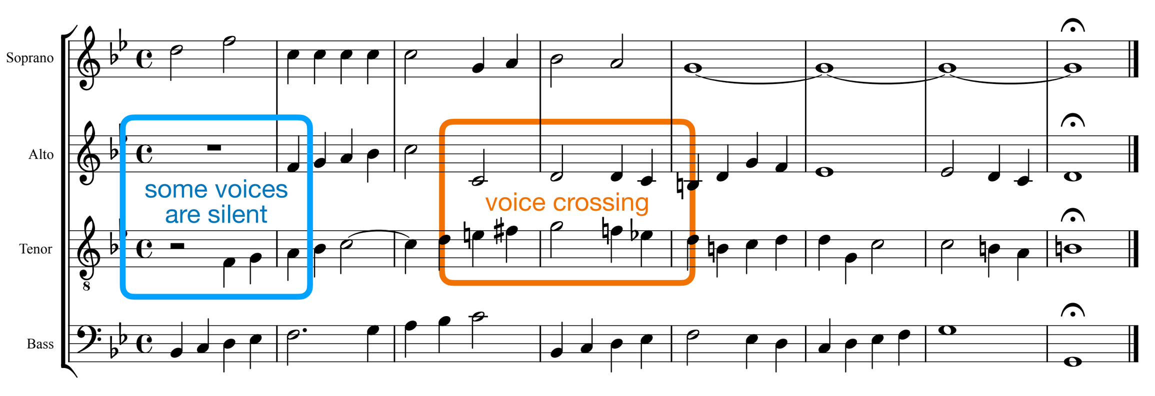


Score-Informed Wave-U-Net

We condition Wave-U-Net on the musical score by feeding a representation of the score into the model as auxiliary input

Why use the score?

Without the score, Wave-U-Net learned to rely on the standard ordering of the voices. Hence, it failed when encountering voice crossings and sections when some voices are silent.



In choral music, sometimes the score may be the only way to associate notes to a specific voice:

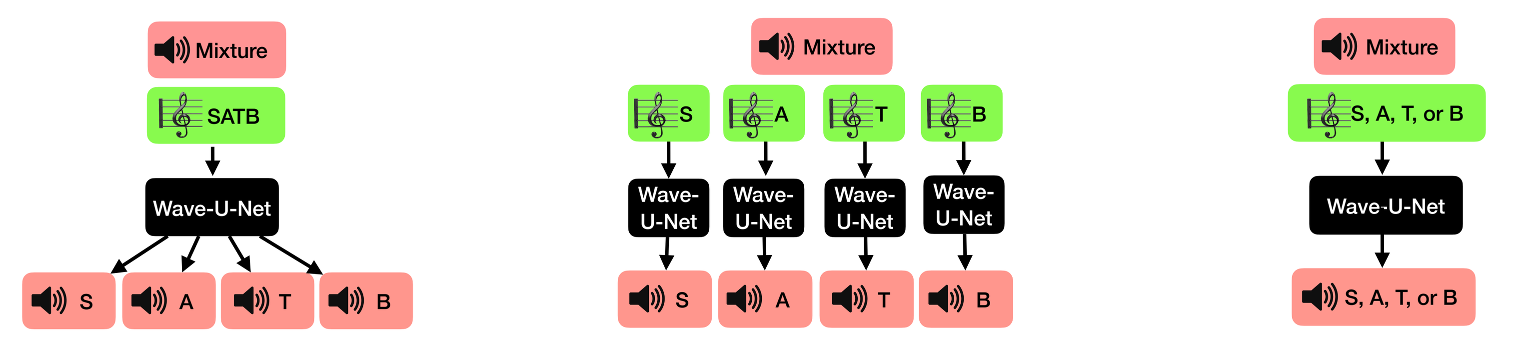
- The timbres of the different choir sections are similar to each other
- Relying on the pitch range is not sufficient because the ranges have considerable overlap
- The standard SATB (soprano-alto-tenor-bass) ordering of the voices is not always kept

Model configurations

One model to extract all sources

Each source is extracted using a separate model

Multi-source model extracts any source (score-guided)



Score representations

We represent a part's score as a time series that indicates the active pitch (if any) at any given time point. We keep the score aligned with the audio by setting the time resolution of the score to be identical to the audio sampling rate.

1. Piano roll

A one-hot matrix of size p x n, where p is the total number of pitches and n is the number of time samples

2. Normalized pitch

A vector containing the active pitch, normalized to the range [0,1]

-1 is used to indicate silence

3. Pitch and amplitude

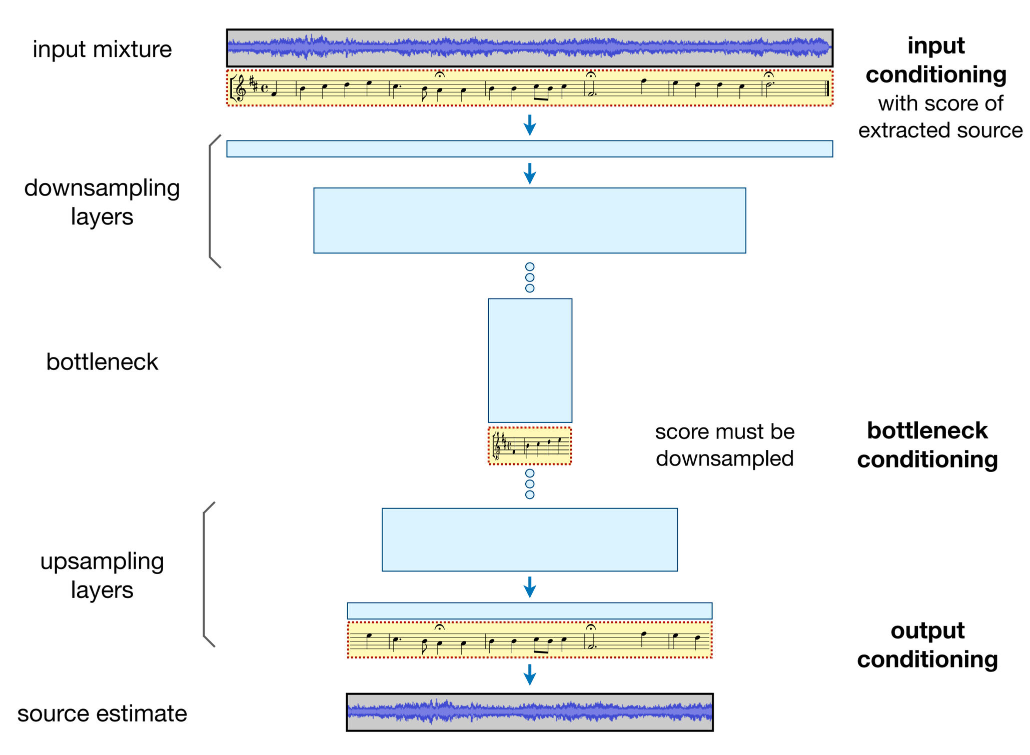
A two-channel representation:

- The pitch channel is a vector containing the active pitch, normalized to [-1,1]
- The amplitude channel contains 1 if any note is active, and 0 otherwise

4. Pure tone

Represents the score in an *audio-like* form: a pure tone signal constructed as a piecewise sine function where the frequency is controlled by the active note's pitch

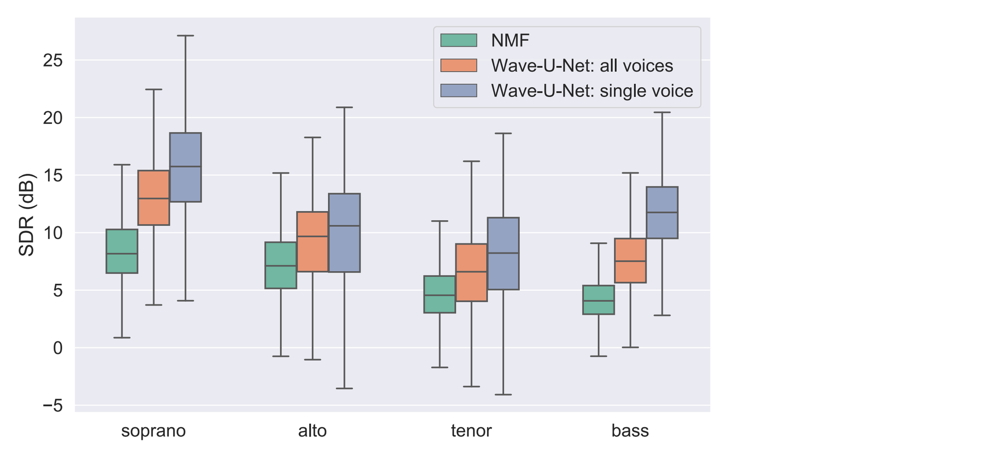
Conditioning locations



Results

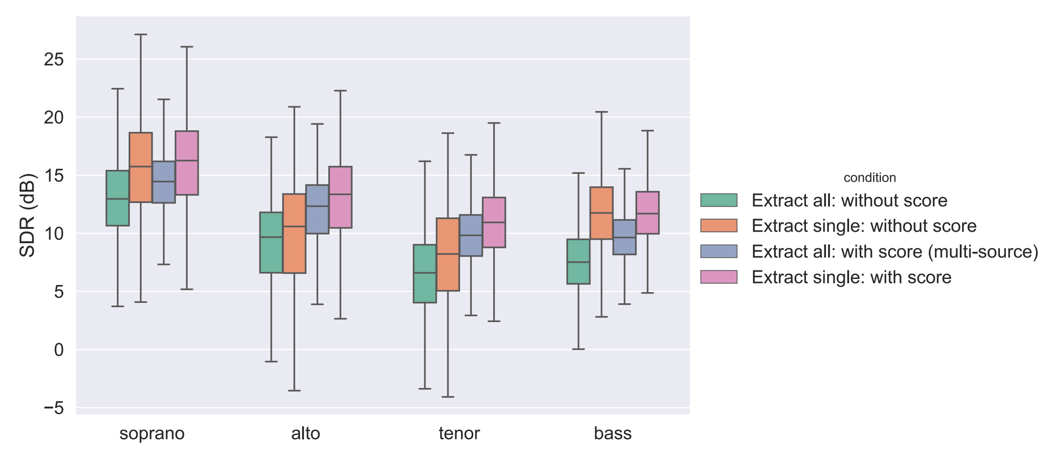
Our main evaluation metric is source to distortion ratio (SDR) provided by the BSS Eval library. [3]

Wave-U-Net vs. NMF



- Wave-U-Net outperforms the NMF baseline by a large margin
- Using a separate model per source performs better than a single model for all sources (but uses 4x the amount of parameters, of course)
- Soprano is easiest to separate. Inner voices are more difficult

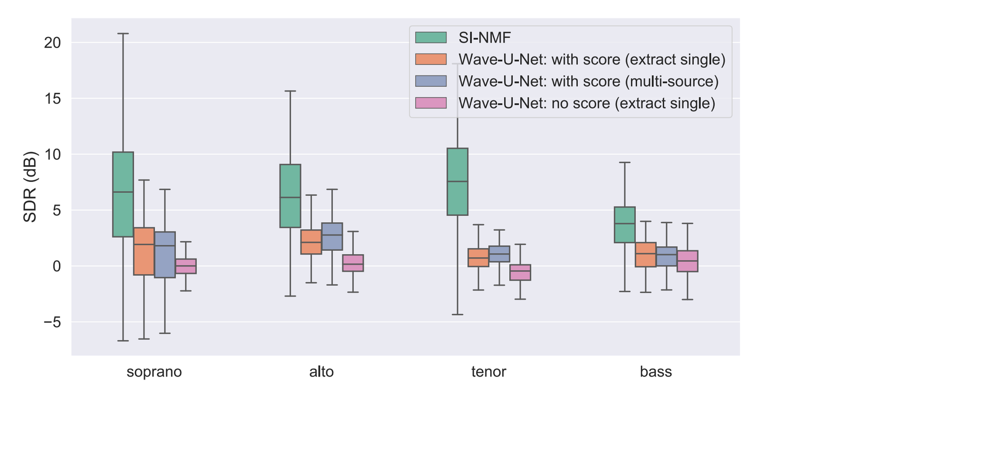
Wave-U-Net: with score vs. without score



- Using the score improves separation performance, especially for the inner voices
- The score is used to disambiguate voice crossings and other difficult cases
- The multi-source (score-guided) model performs well even though it uses only a single model to extract any of the sources

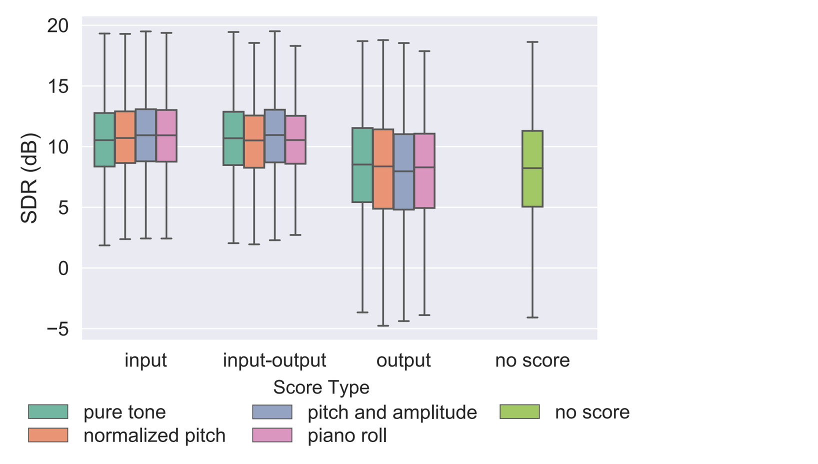
Evaluation on real choir recordings

Using recordings from Choral Singing Dataset [4]



- Wave-U-Net trained on our synthesized dataset does not generalize well to real recordings. Score-informed NMF still performs better in this case
- Future work to make the model more robust to real recordings:
  - Create and use a training dataset of real recordings
  - Use more advanced singing synthesis methods to incorporate lyrics
  - Use better data augmentation techniques

Comparing score conditioning methods



- The choice of score representation does not make a big difference
- We notice that score conditioning leads to artifacts at note boundaries (likely due to the discontinuity of the score representations). The pure tone score type reduces these artifacts
- Conditioning only at the output layer performs badly, likely because the output layer is merely a dot product (convolution with kernel of size 1). Models conditioned only at the output have learned to ignore the score
- Future work could try more complex conditioning methods such as FiLM [5]

References

[1] S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in *ICASSP* 2012

[2] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *ISMIR* 2018

[3] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *LVA/ICA* 2018

[4] H. Cuesta, E. Gómez, A. Martorell, and F. Loáiciga, "Analysis of intonation in unison choir singing," in *ICMPC* 2018

[5] G. Meseguer-Brocal and G. Peeters, "Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separations," in *ISMIR* 2019