Score-Informed Source Separation of Choral Music

Matan Gover and Philippe Depalle
McGill University

http://www.matangover.com/choirsep

https://program.ismir2020.net/poster_2-09.html

Presented at 21st ISMIR Conference, October 2020

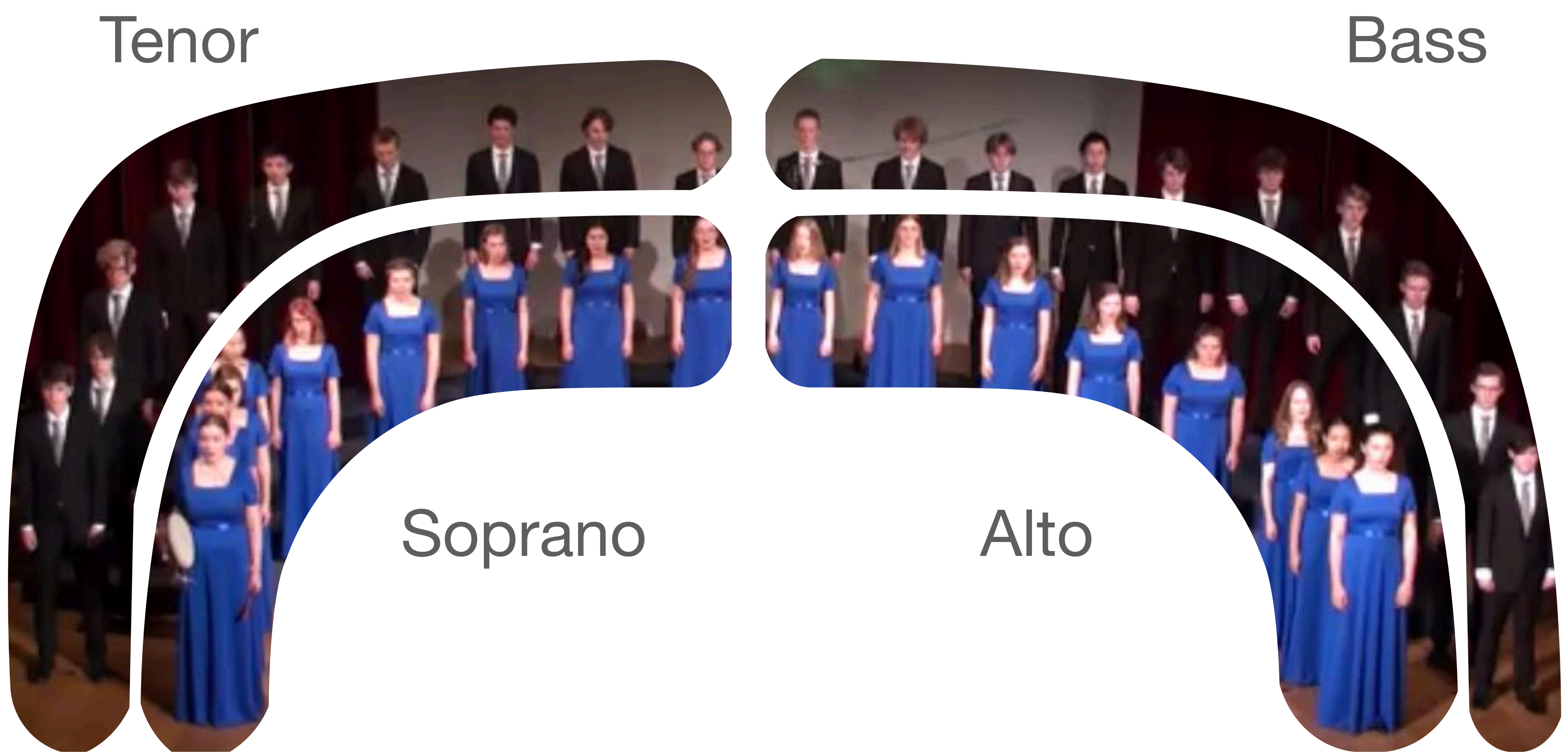
McGill Schulich School of Music
École de musique Schulich

Our Task

Source separation of choral music

input: choir recording

output: individual track for each choir section



Motivation

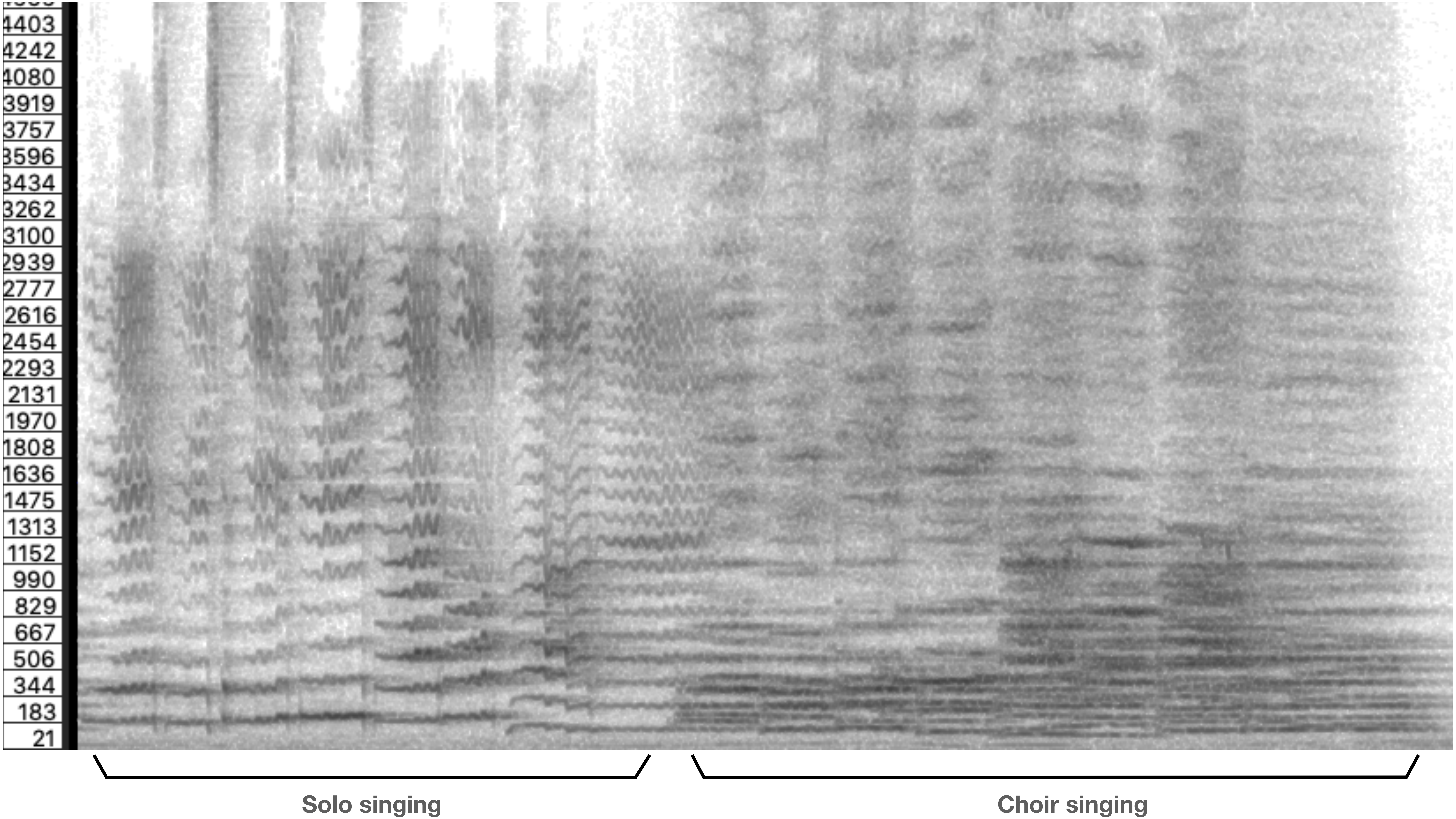
- Personal interest in choir music
- New task, no baseline to compare to
- Fine-grained editing, mixing, and analysis
- Automatic creation of choir practice tracks


Master Your Choral Music & Have a Great Pe
SingleParts™ with StudySpots™
"In Pursuit of Choral Excellence"
Home
About
Warm-up
** Now available... IMMEDIATELY
Hear Samples
What Is SingleParts
How SingleParts Is Made
Conductor Comments
Singer Comments
Other Helpful Websites
Choral Practice Tracks on CD and M
• Automatic crea
CHORALPRACTICE
ONLINE VOICES
HOME
WORKS
RADIO
FAQ
LATEST NEWS
CONTACT
online application!
the Lord God om - ni - po - tent reign - eth,
Rehearsal Aids
Helping Choirs Learn, Rehearse and Practice Choral Music
Our Mission
Rehearsal Aids ma
learn and rehearse
We work with choi
Choralia
ing aids downloaded so
809
files are provided, which allow singers to gradually
re also "sung" by electronically-synthesized human
S), simulated voices using a simple "doo doo" sound,
and it is now discontinued (although all training aids produced in CMS technology remain available).
Choral Tracks
Practice Made Simple
Browse
Whom We Serve
About Us
Free Online Learning
All State
Search
Become A Member
Login
Can't Rehearse Together?
Get the BEST Online Learning Solution
for Full Choirs - Only $999.99/Year
Try Us For 7 Days Risk Free - Sign-Up Below!
Individual Memberships and Digital Downloads Available Too

Challenges

Why is choral music hard to separate?

- Each section is actually multiple singers with varying pitch, timbre and timing
- Separation must “undo” choral blend
- Lack of datasets


4403
4242
4080
3919
3757
3596
3434
3262
3100
2939
2777
2616
2454
2293
2131
1970
1808
1636
1475
1313
1152
990
829
667
506
344
183
21
Solo singing
Choir singing

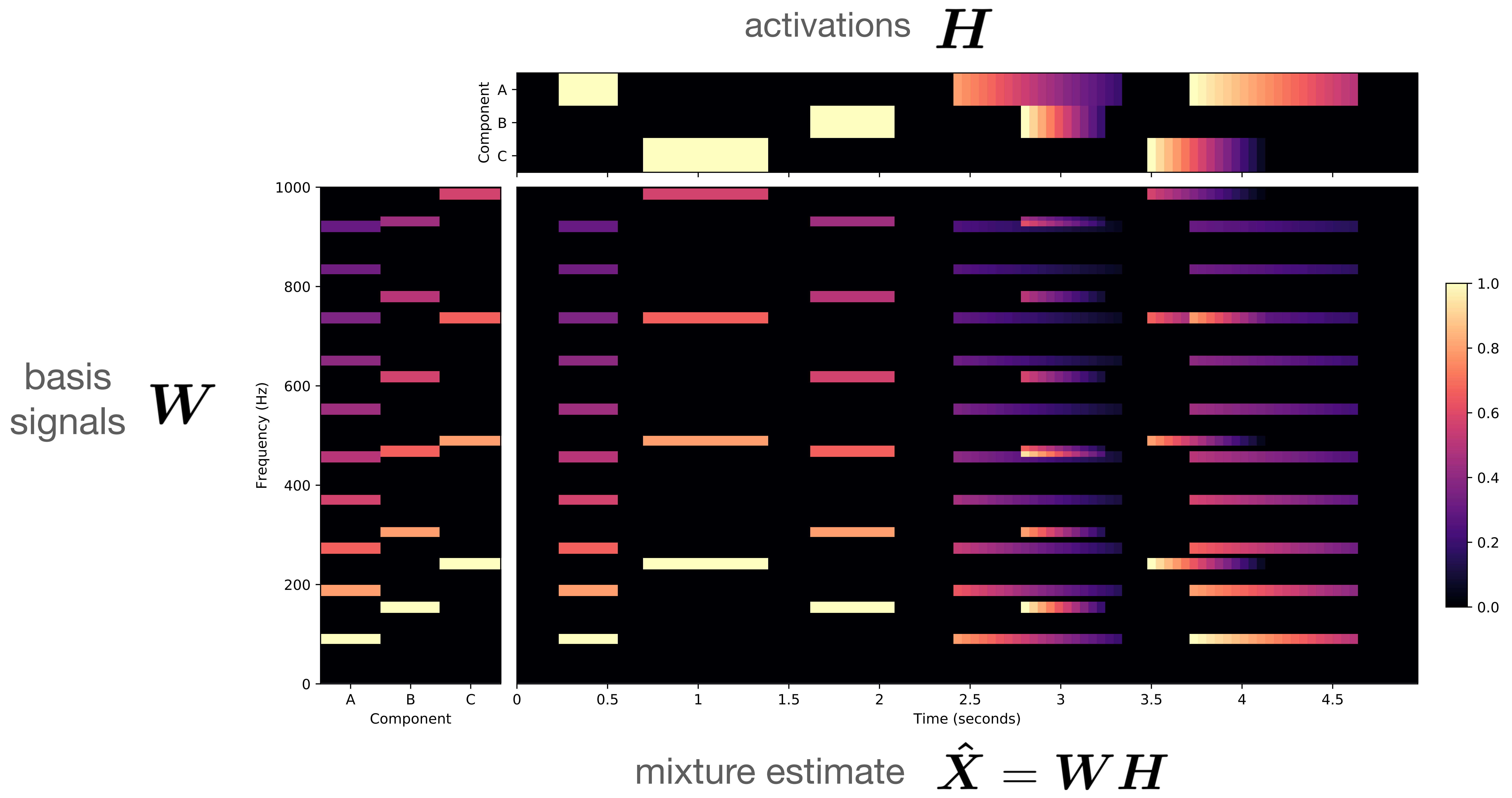
Methods

- Unsupervised — **Score-informed NMF**
- Supervised, with synthesized dataset

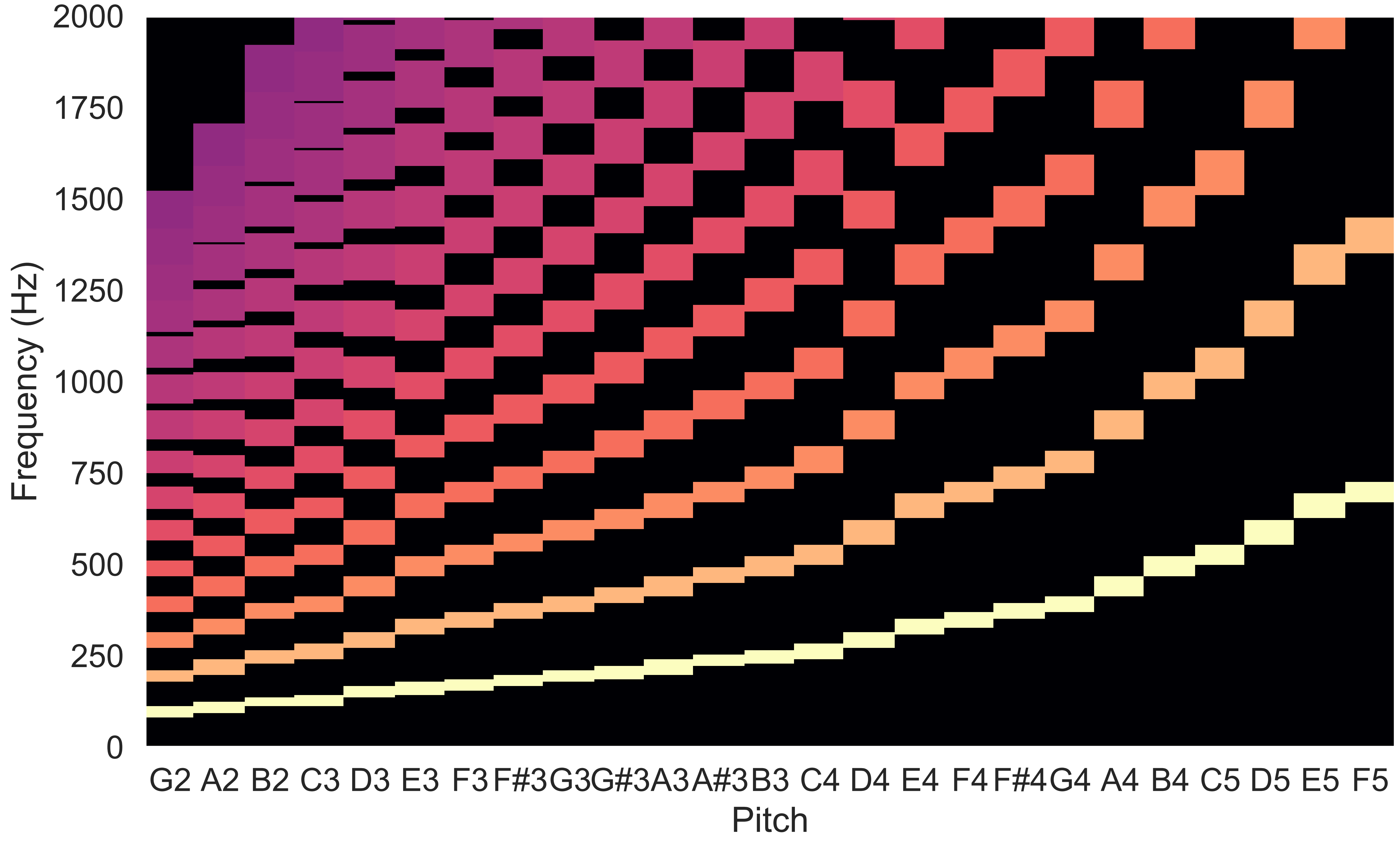
Baseline: Score-informed NMF

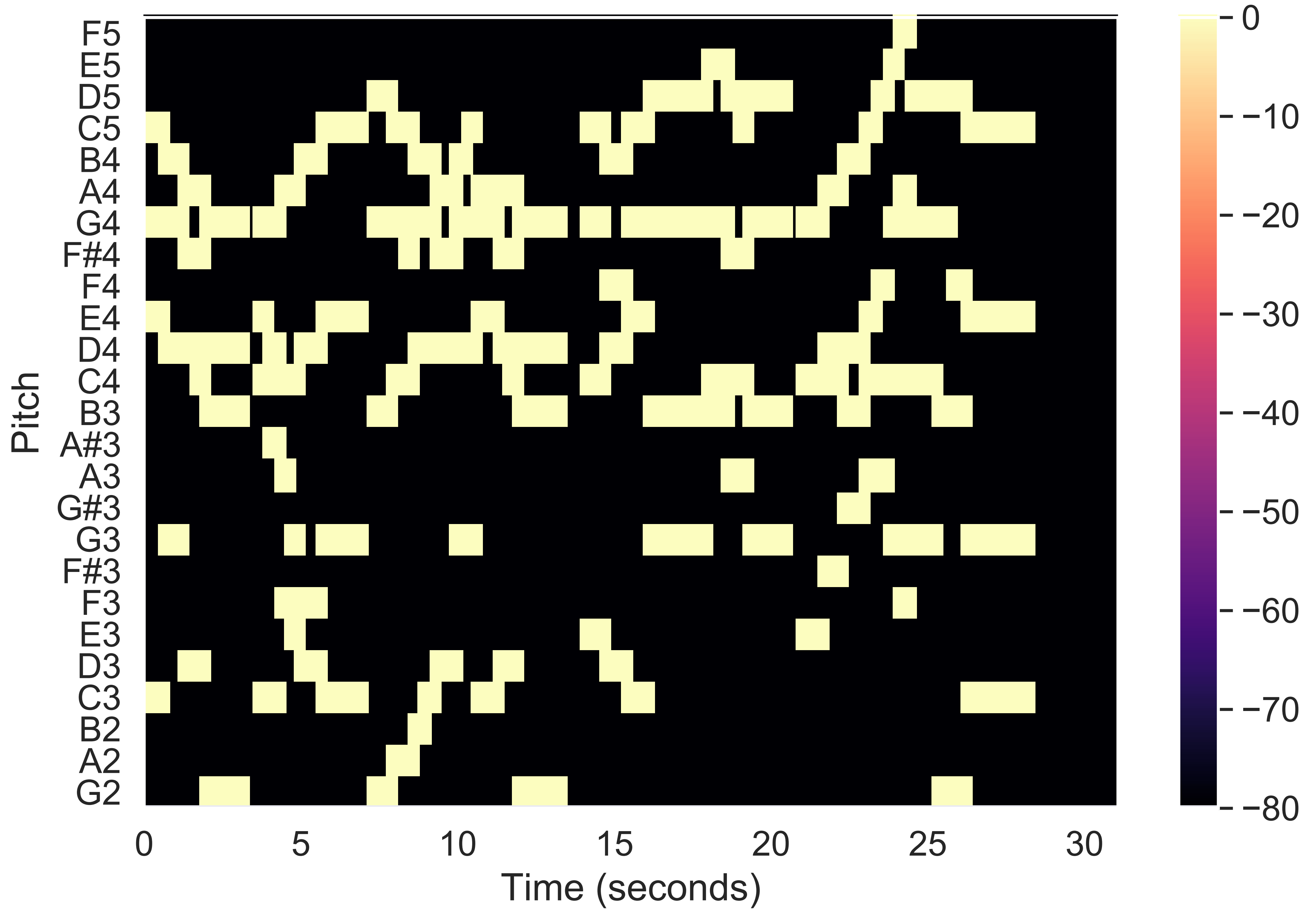
- Factorizes mixture spectrogram as a product of two matrices: basis signals and activations
- Ratio mask is applied to the mixture spectrogram for extracting each source
- Constrained using timing and pitch information from score using a technique originally used for piano notes [Ewert and Müller, 2014]

Factorizes mixture spectrogram as a product of two matrices:



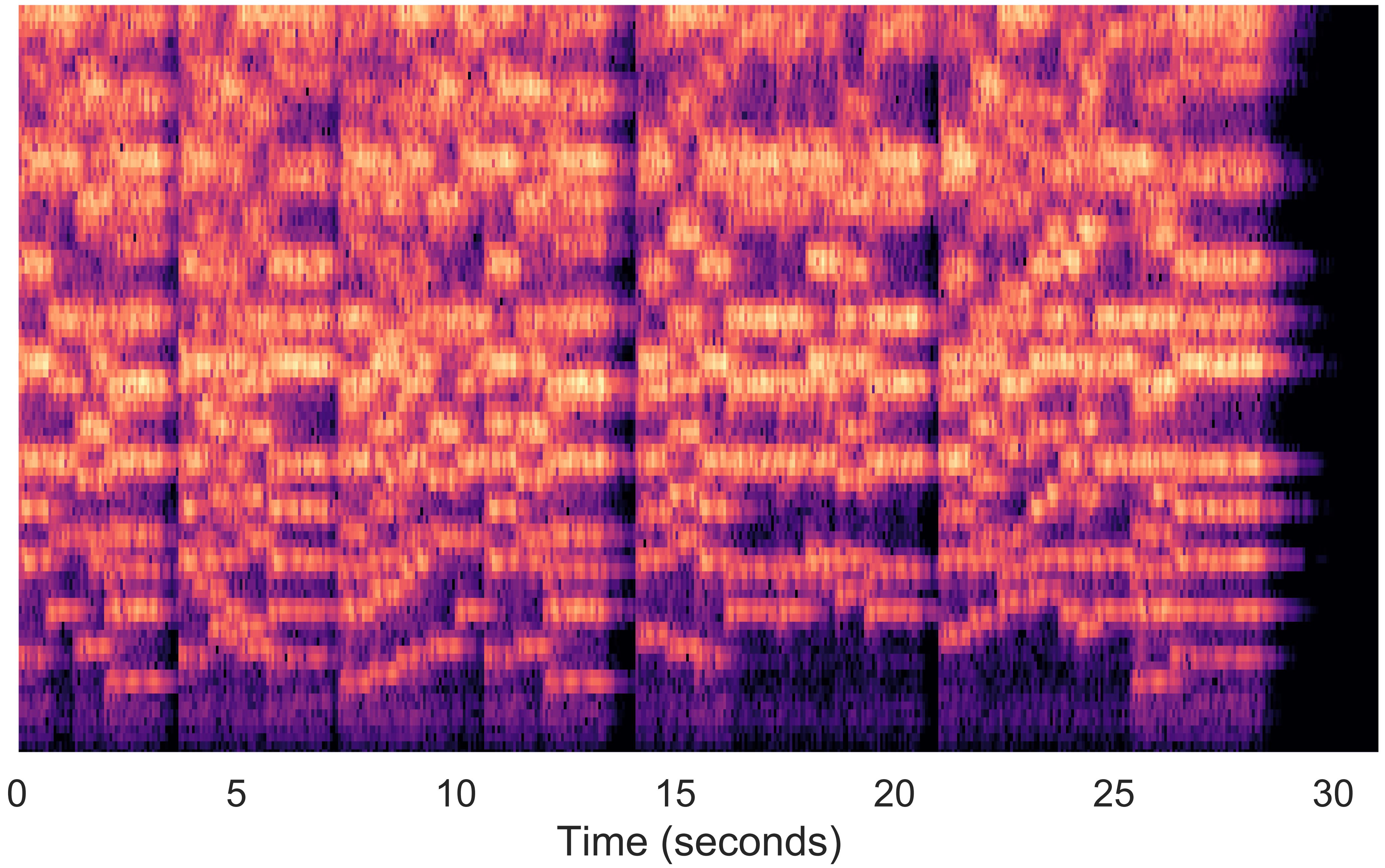
NMF initializations are constrained using timing and pitch information from the musical score:





https://github.com/matangover/score-informed-nmf

Score-informed NMF



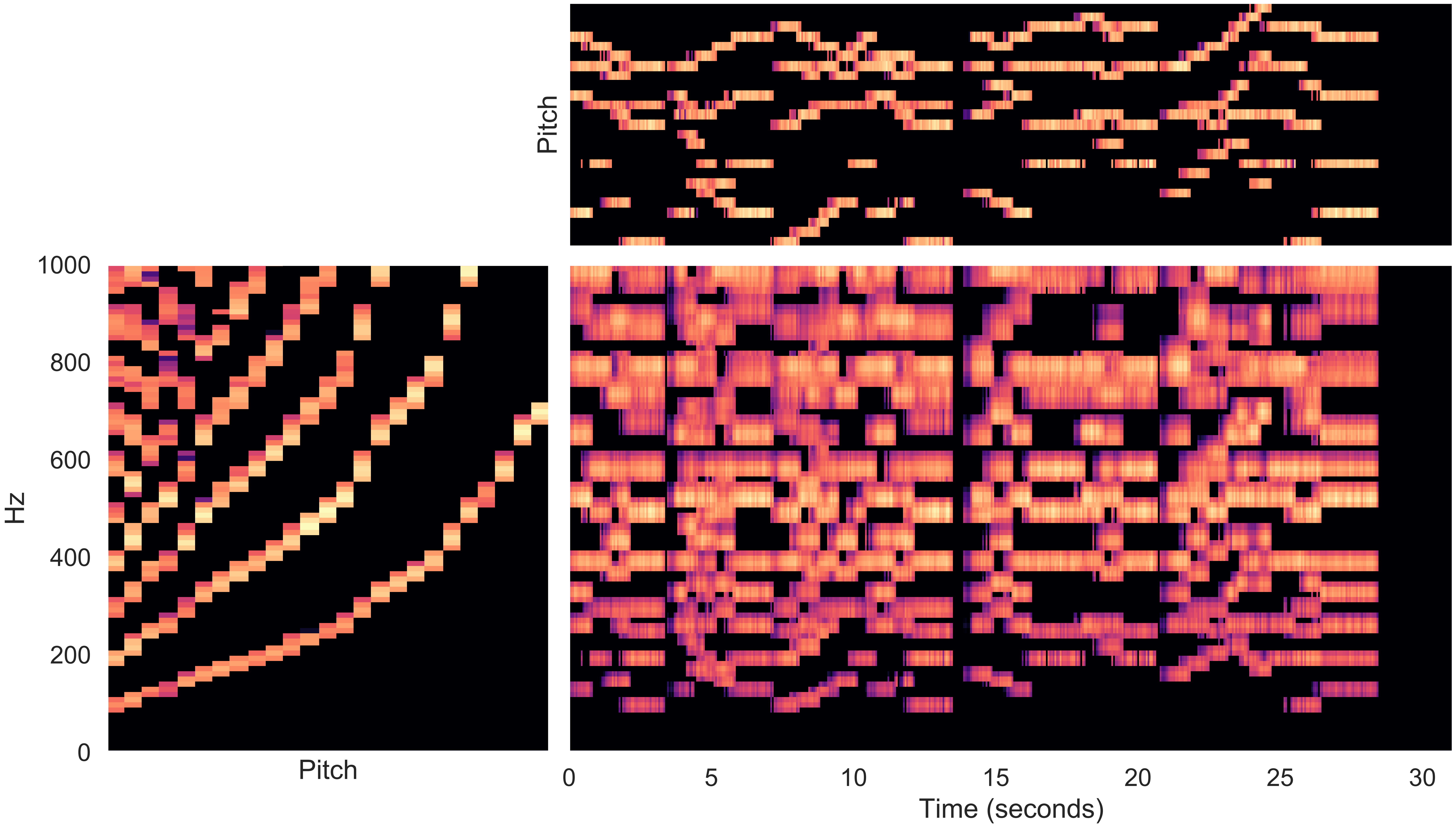
mixture

$\boldsymbol{X}$

Score-informed NMF

activations H

basis signals W



mixture (approximate)

$$\hat{X} = WH$$

Score-informed NMF – Results

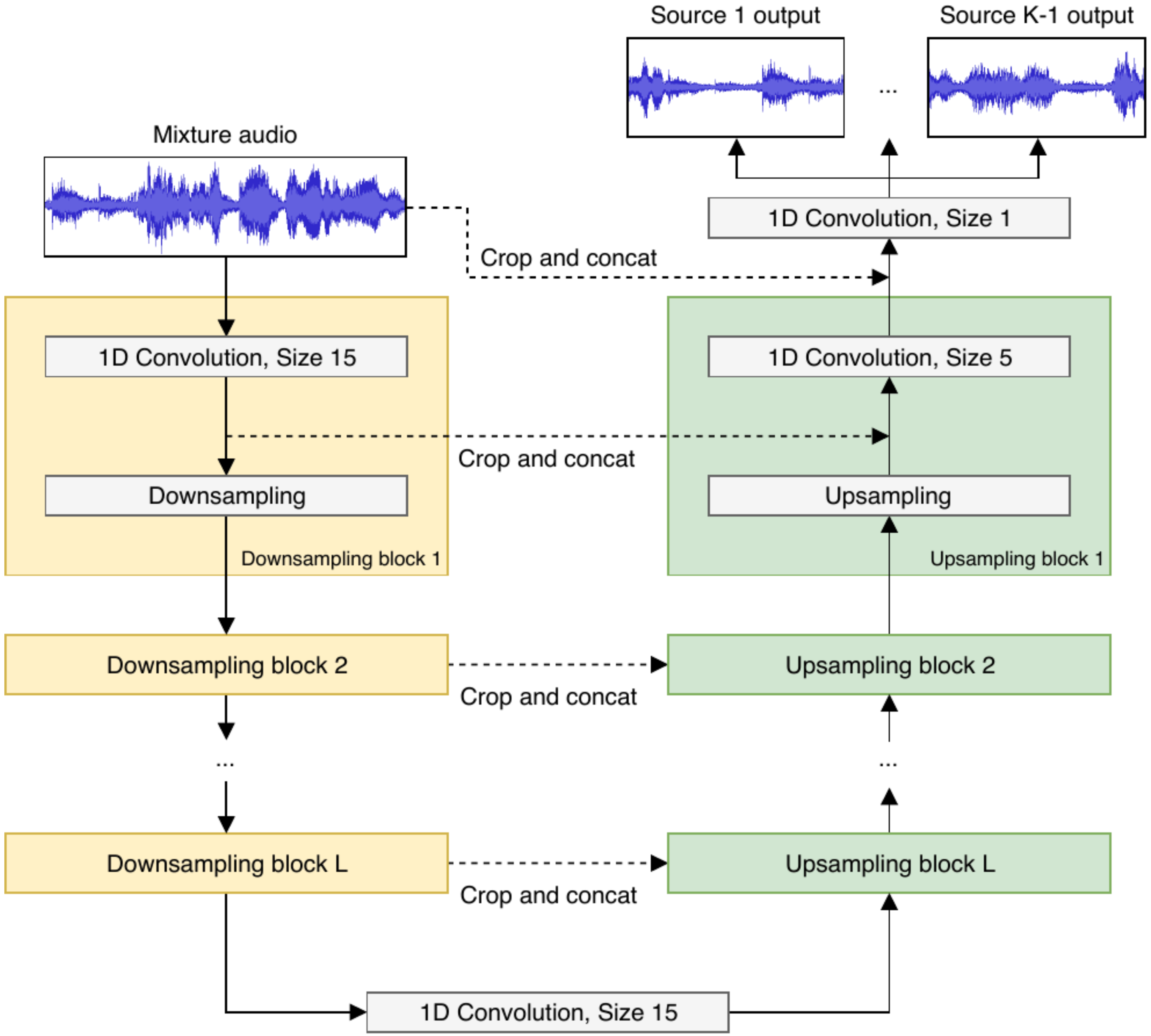
- Does not capture continuous evolution of pitch and timbre
- Undesirable amplitude fluctuation artifacts

Methods

- Unsupervised **Score-informed NMF**
- Supervised, with synthesized dataset **Score-informed Wave-U-Net**

Wave-U-Net

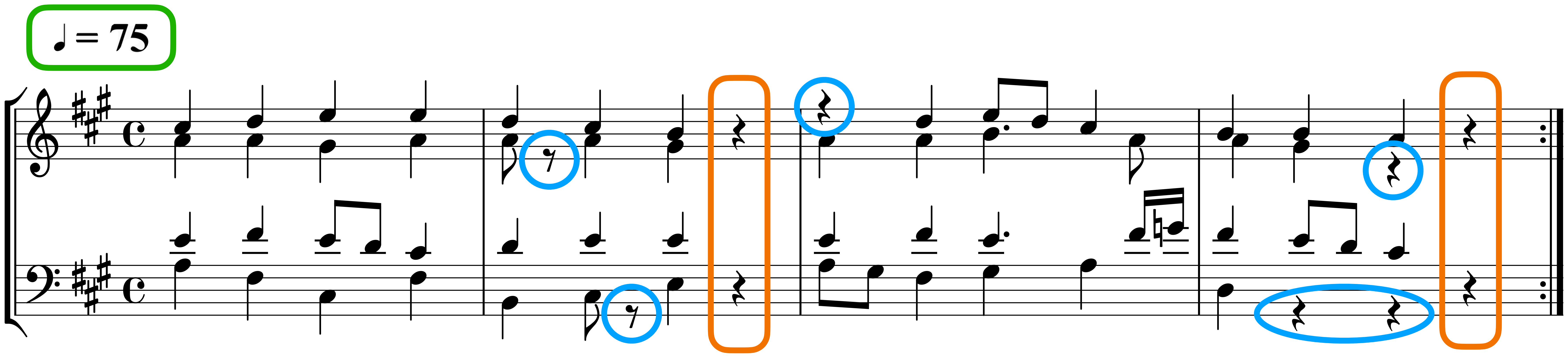
- Encoder-decoder with skip connections
- Worked well for vocals & accompaniment separation
- Works directly on the time-domain signal



Synthesized Choir Dataset

Bach chorale harmonizations

- 351 chorales (~4 hours)
- Sample-based synthesis (no lyrics)
- Data augmentation: simulated breaths, random omitted notes, and tempo variations

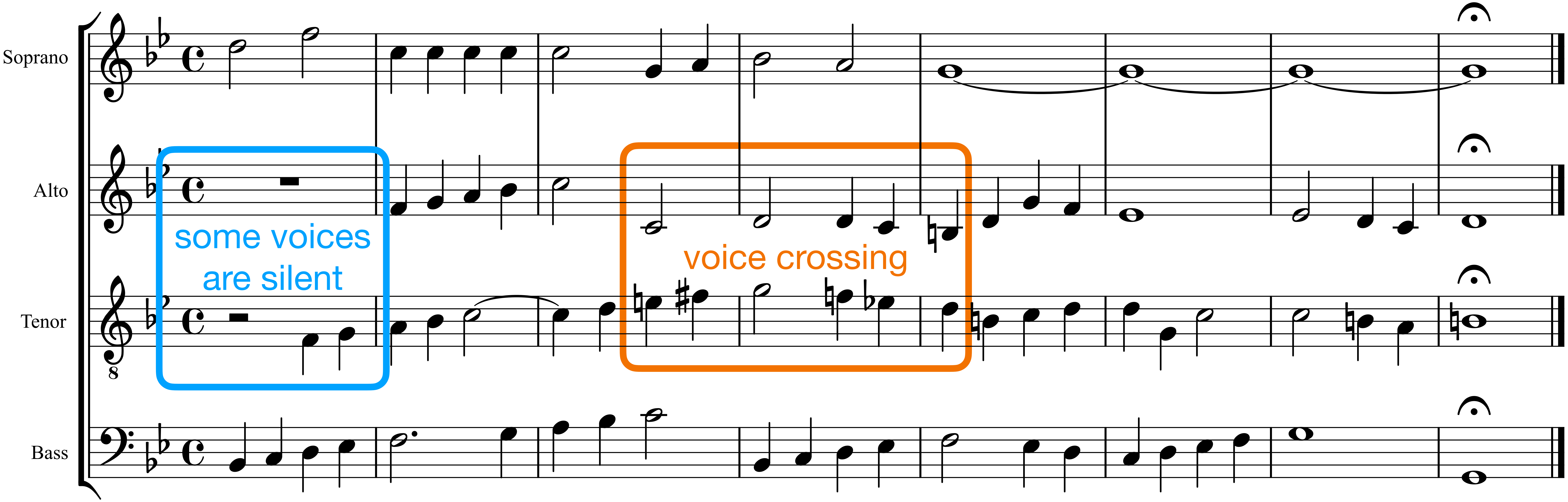


Example: chorale BWV 359 original, augmented, synthesized

https://github.com/matangover/synthesize-chorales

Problem: voice crossings

Model learned to rely on SATB ordering of voices



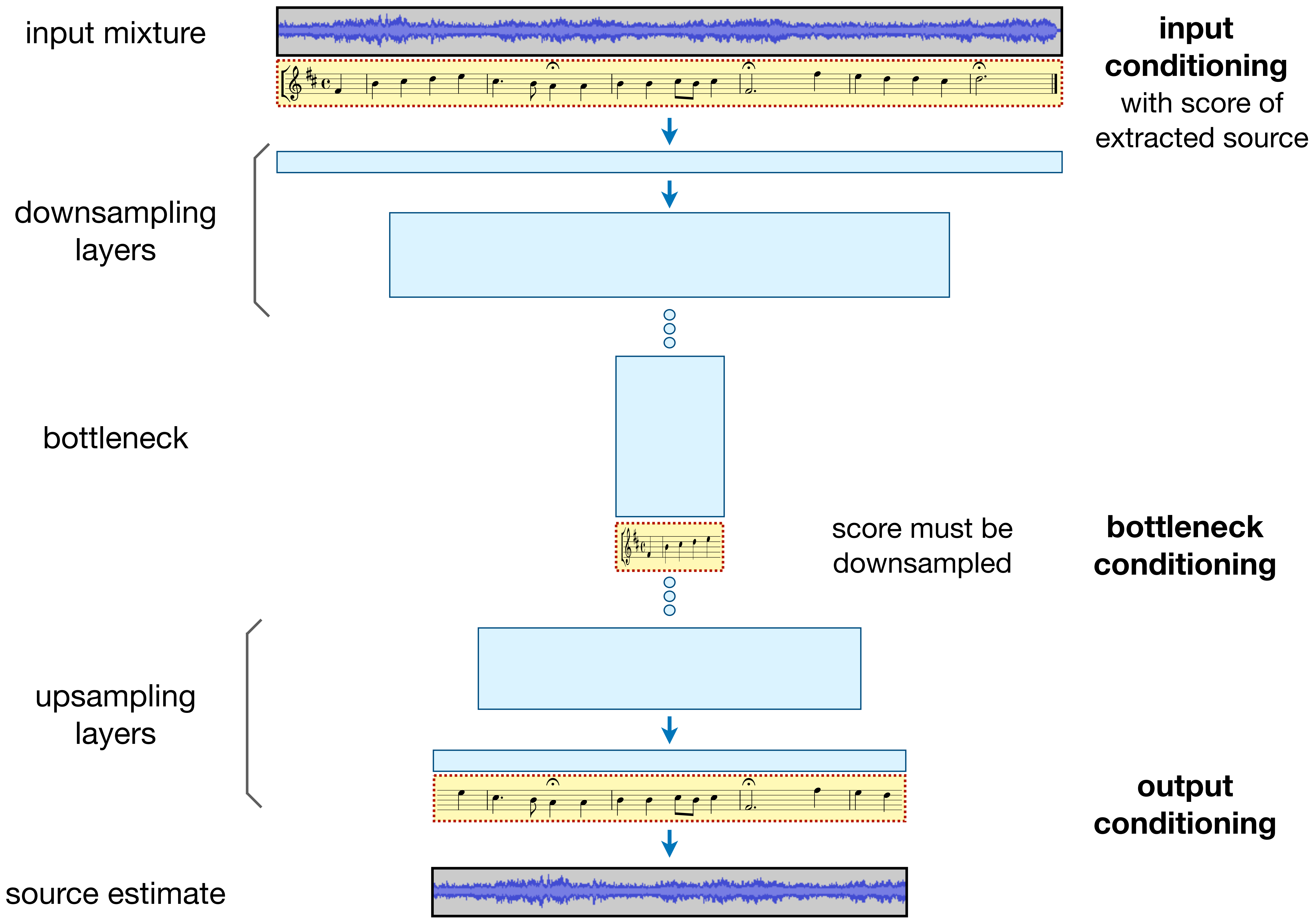
Conditioning on Score

- Part's score represented as a time series: indicates the active pitch (if any) at any given time point
- Score aligned with the audio: score time resolution is identical to audio sampling rate.
- 4 score representations x 3 conditioning locations

Score Representations

- <u>Piano roll</u>: 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
- <u>Normalized pitch</u>: A vector containing the active pitch, normalized to the range [0,1]. -1 is used to indicate silence
- <u>Pitch and amplitude</u>: 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
- <u>Pure tone</u>: 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

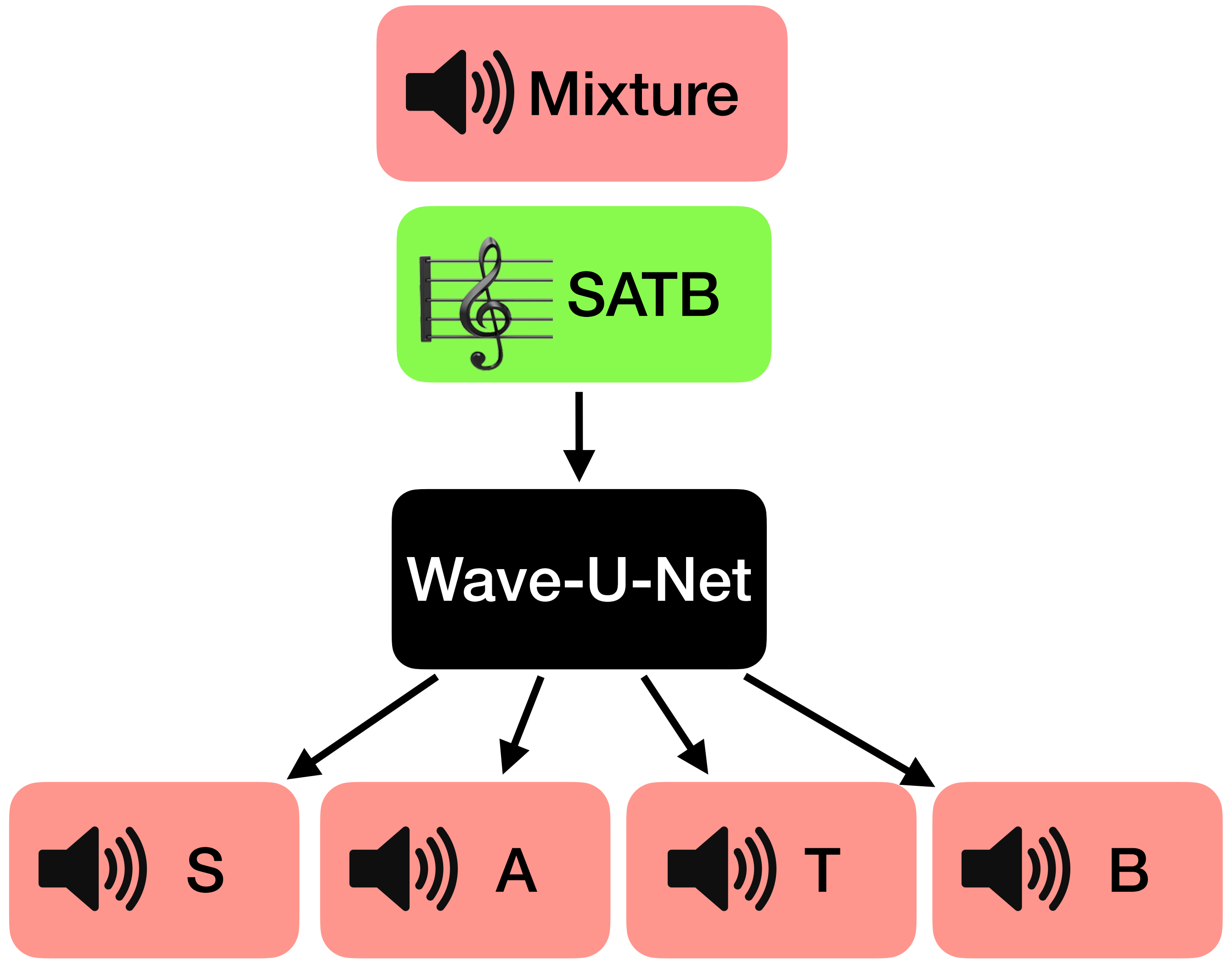
Conditioning Locations



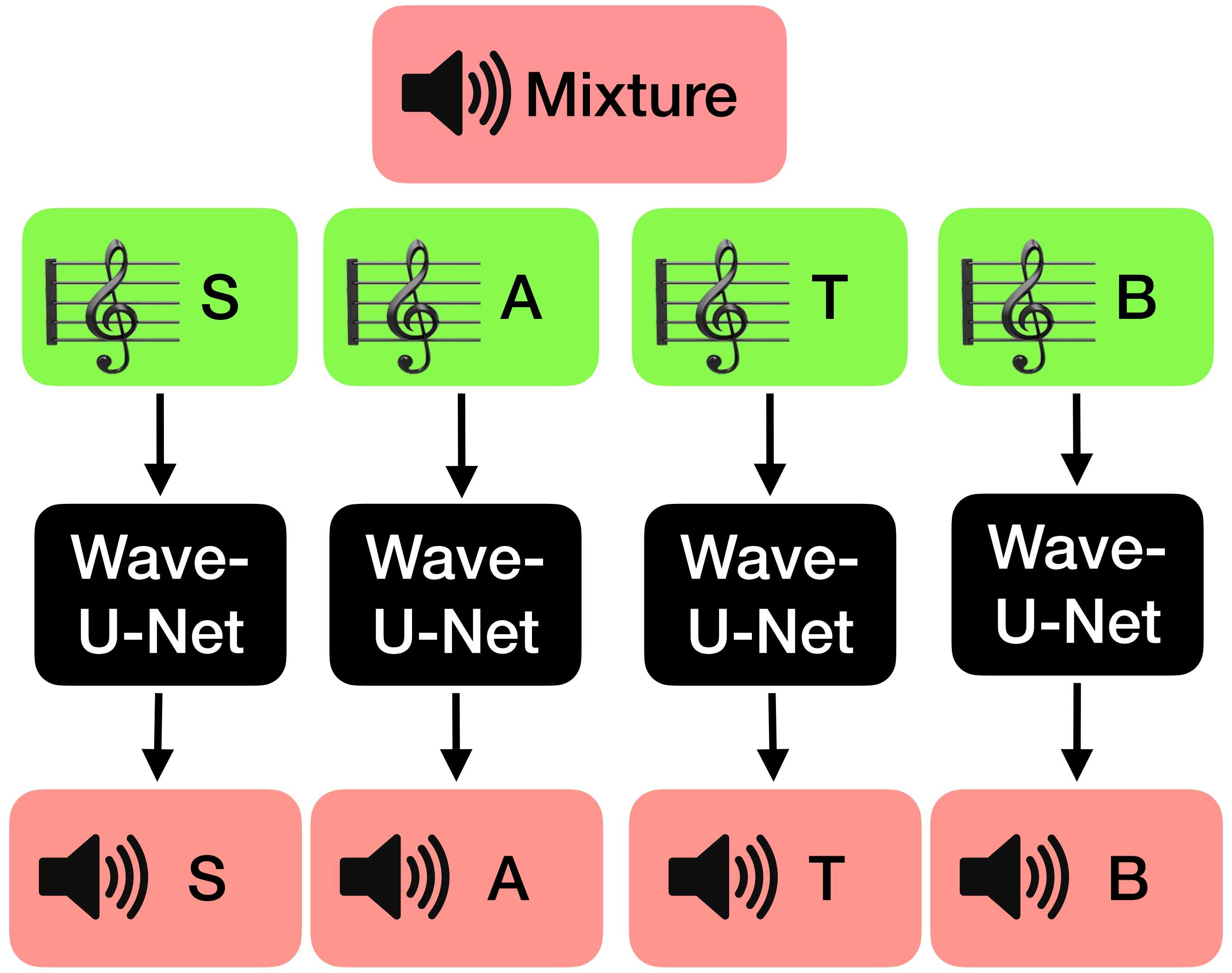
Score-informed Wave-U-Net

Model configurations

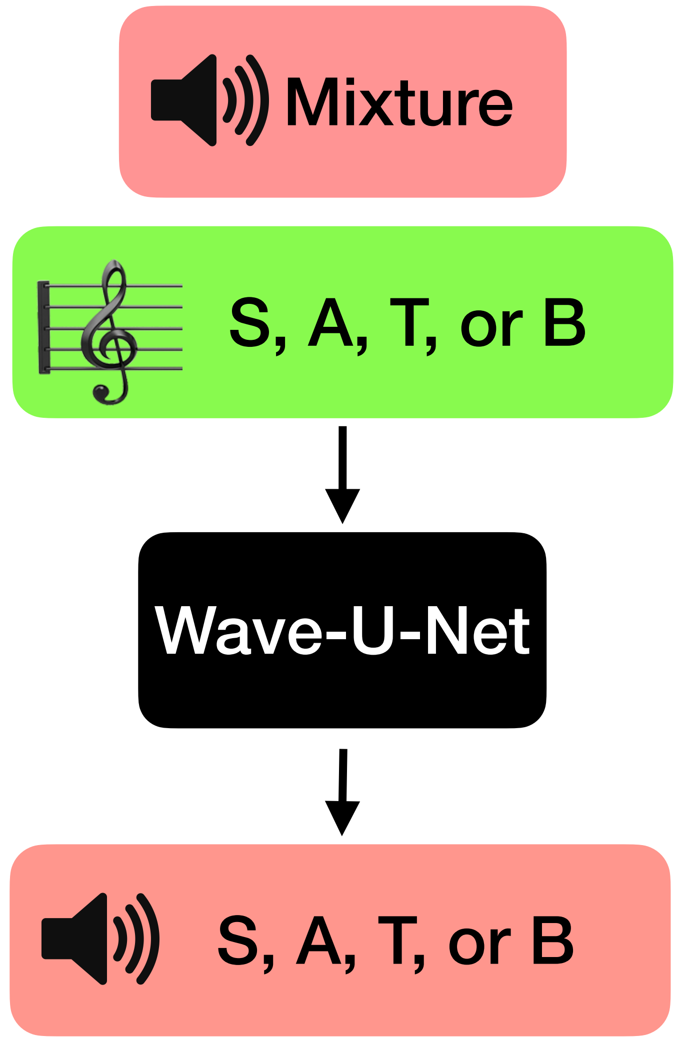
One model to
extract all sources



Each source is extracted
using a separate model



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

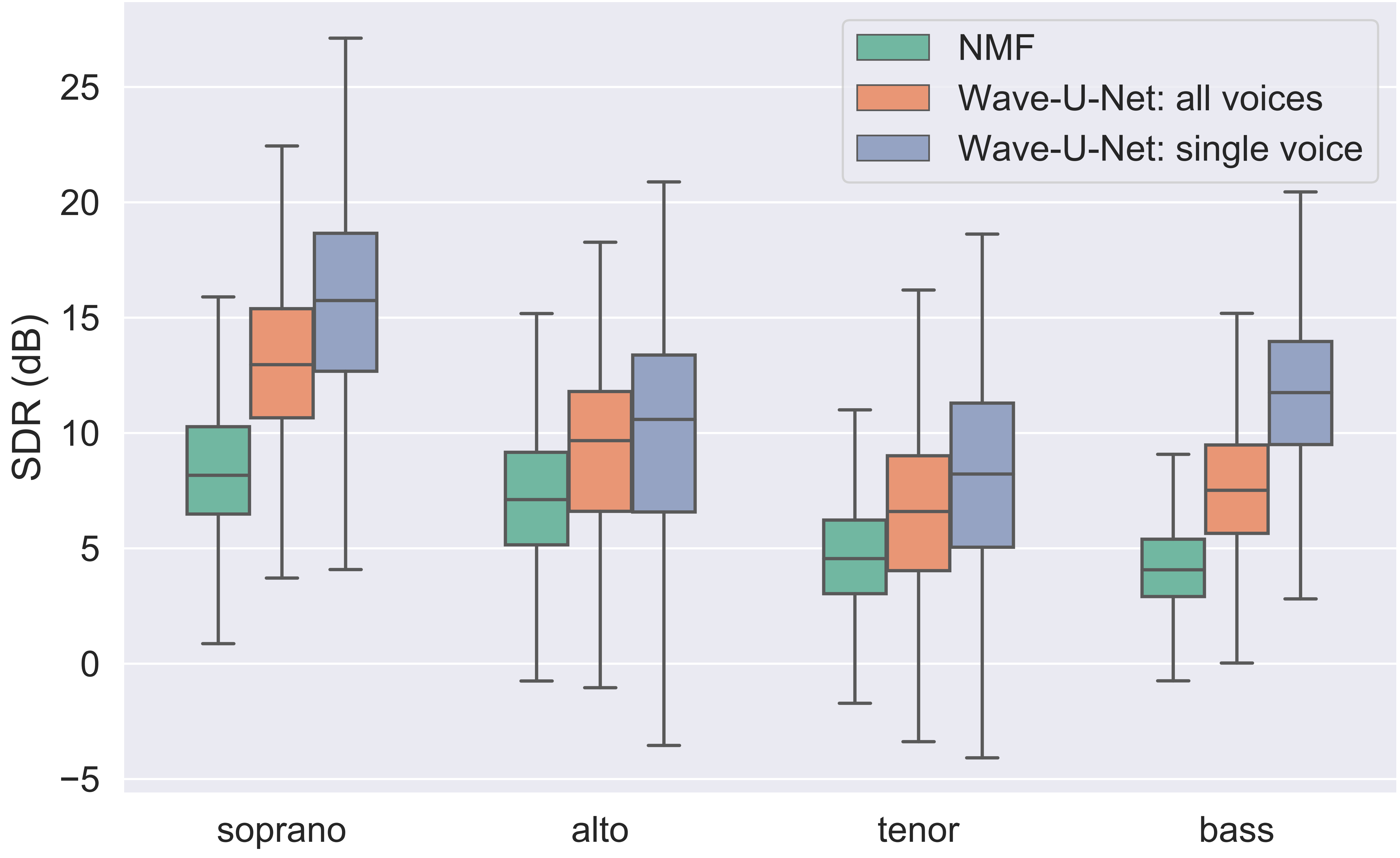


Experiments

Experiment	Method	Score-Informed	Model Type
1	SI-NMF	yes	-
2	Wave-U-Net	no	one model for all voices
3	Wave-U-Net	no	one model per voice
4	Wave-U-Net	yes	one model for all voices
5	Wave-U-Net	yes	one model per voice
6	Wave-U-Net	yes	one model: multi-source

Results

Wave-U-Net vs. NMF

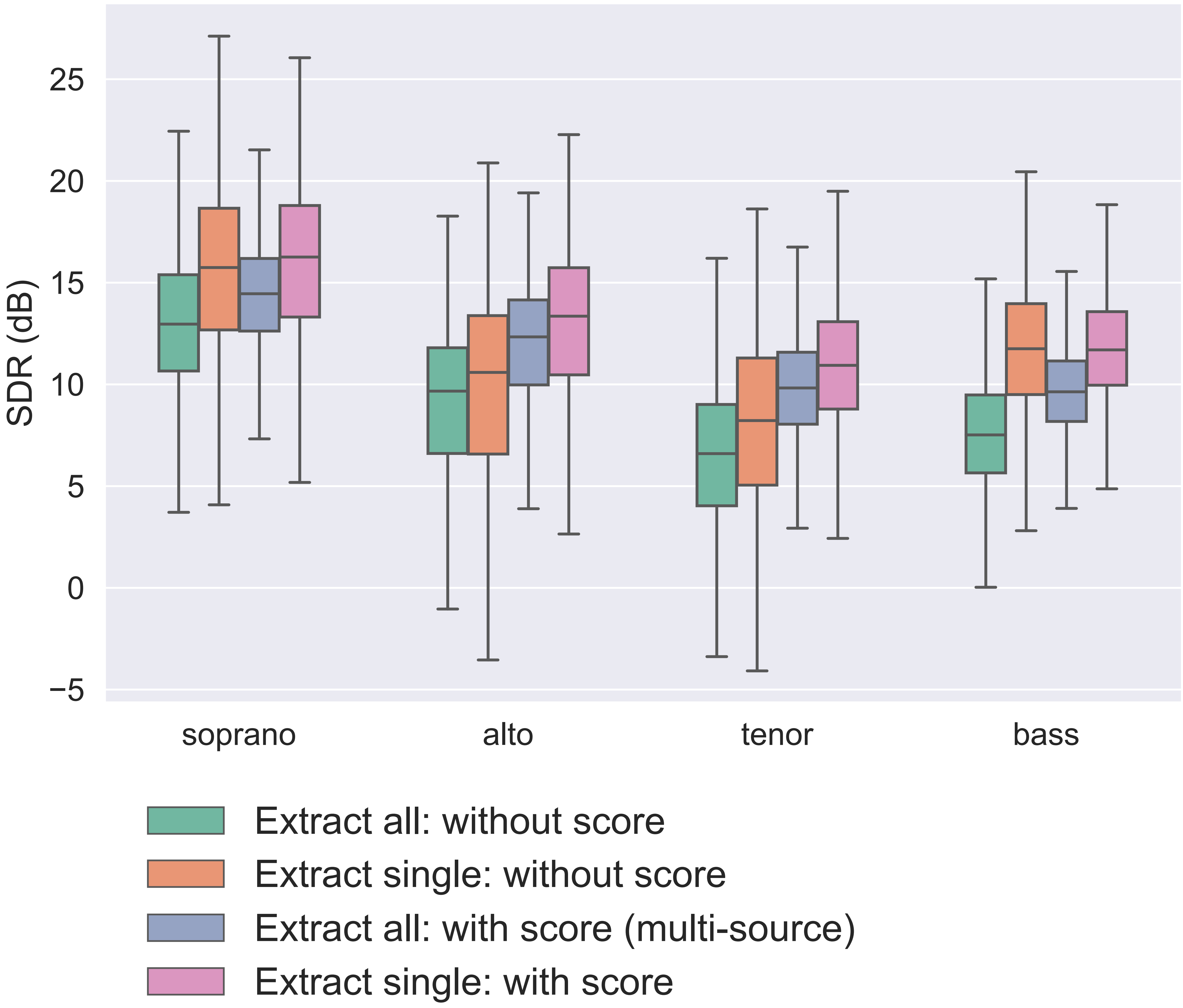


Evaluation metric: source to distortion ratio (SDR) provided by the BSS Eval library.

- 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

Results

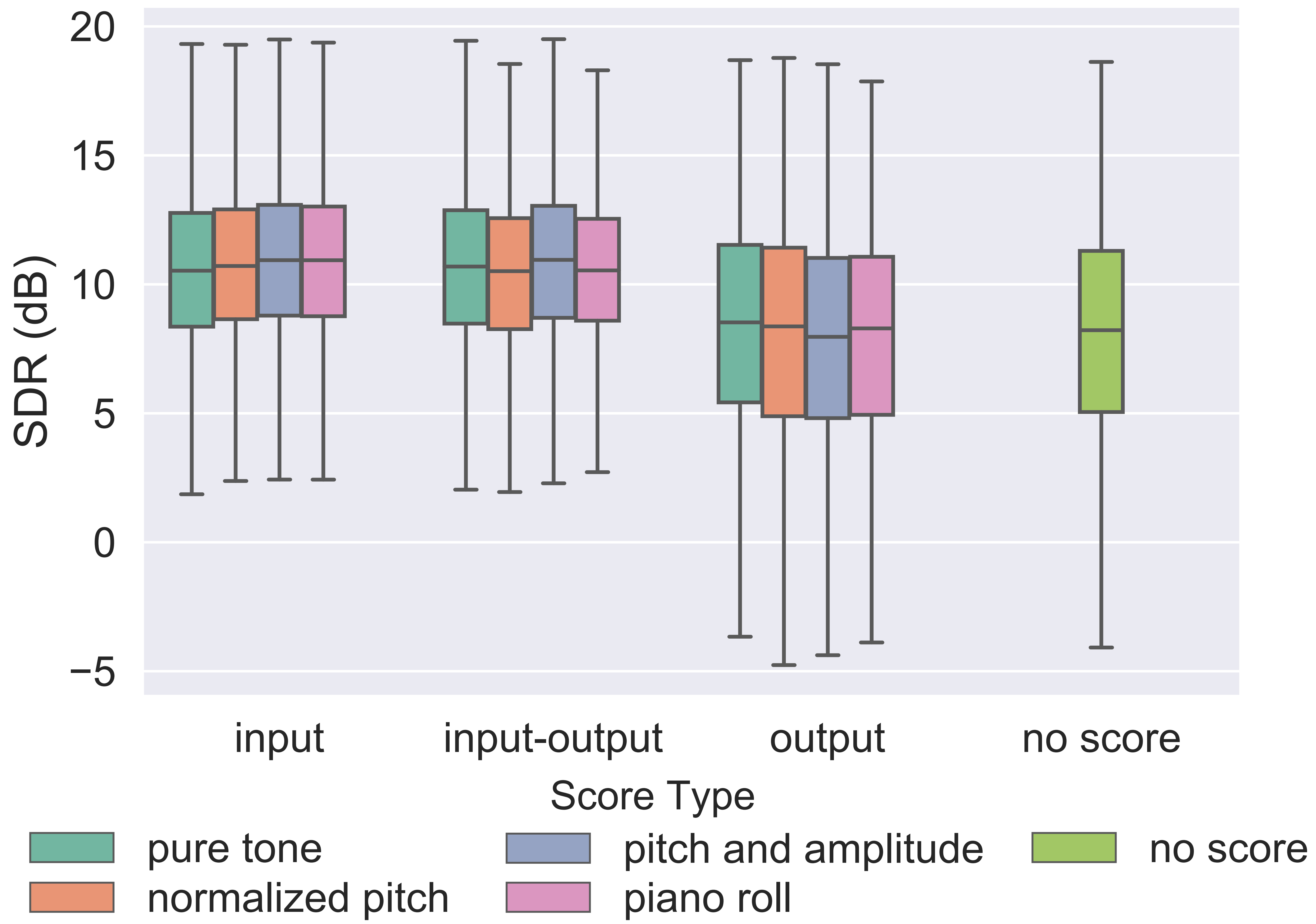
Wave-U-Net: with score vs. without score



- 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

Results

Comparing score conditioning methods

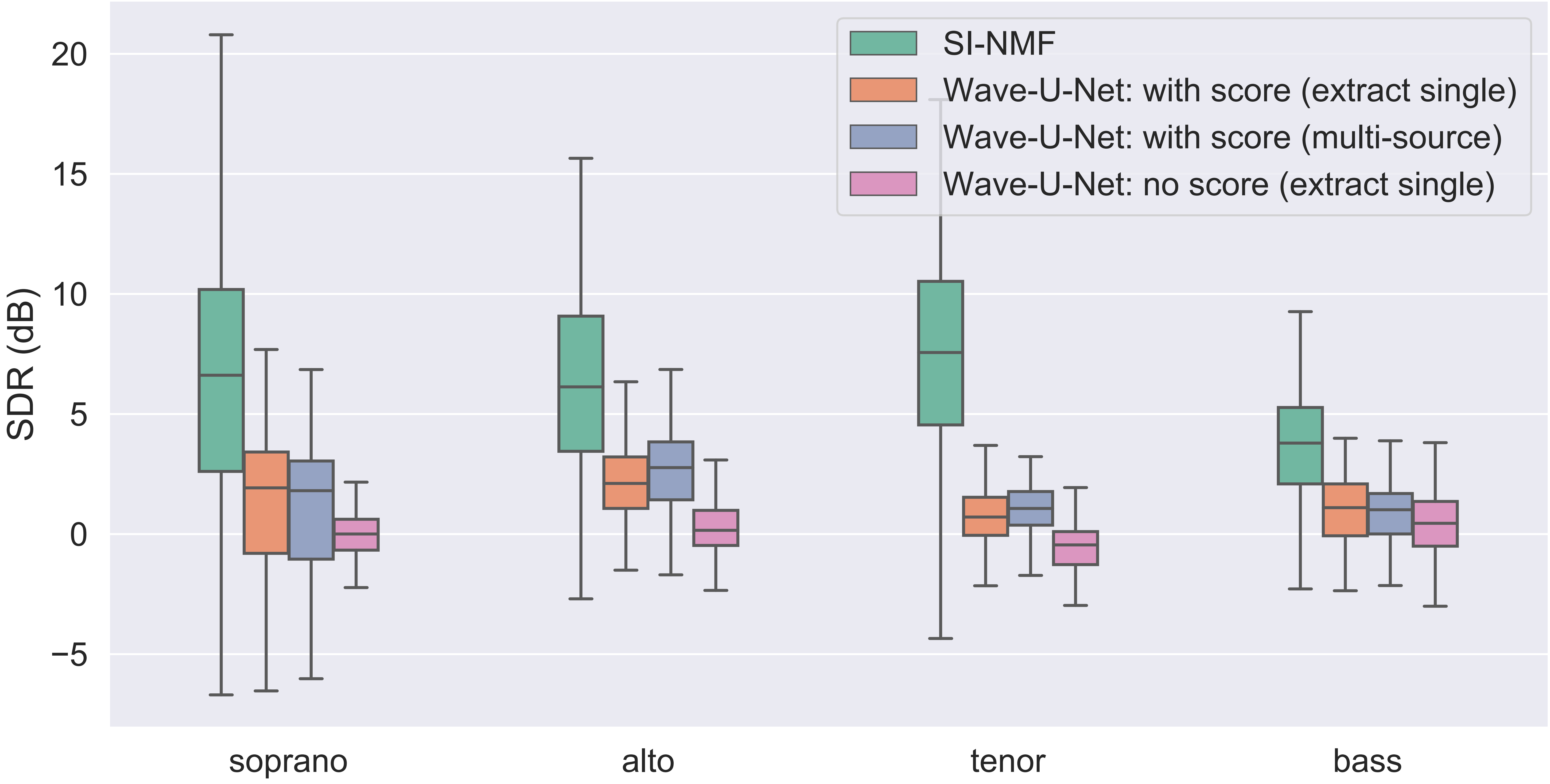


- Score representation does not make big difference
- Score conditioning leads to artifacts at note boundaries
  - ✳ due to the discontinuity of score?
  - ✳ Pure tone score type reduces these artifacts
- Conditioning at the output layer performs badly
  - ✳ likely because the output layer is merely a dot product
- Try more versatile conditioning, e.g. FiLM

Results

Evaluation on real choir recordings

Using recordings from Choral Singing Dataset



Wave-U-Net trained on synthesized dataset does not generalize well to real recordings.

Score-informed NMF still performs better in this case

Results – Bottom Line

- Wave-U-Net outperforms NMF on synthesized dataset by large margin
- Score is successfully used to disambiguate misclassified notes
- NMF still performs better on real choir recordings

Next steps

- Still some way to go for real choir music [people are working on it]
  - Need multi-track choir datasets:
    - Collaborate with learning track websites, virtual choir initiatives
    - Better synthesis methods: automating choir VSTs, using modern choir synthesis
  - Unsupervised and semi-supervised: Mixtures of mixtures
- Integrate instrumental accompaniment separation
- Non-aligned scores (joint 'transcription' and separation)
- Input features: Spectrograms or learned filter banks

Thank you!

https://www.matangover.com/choirsep-ismir