Conference Review

Katharina Prinz, Sebastian Böck, OFAI

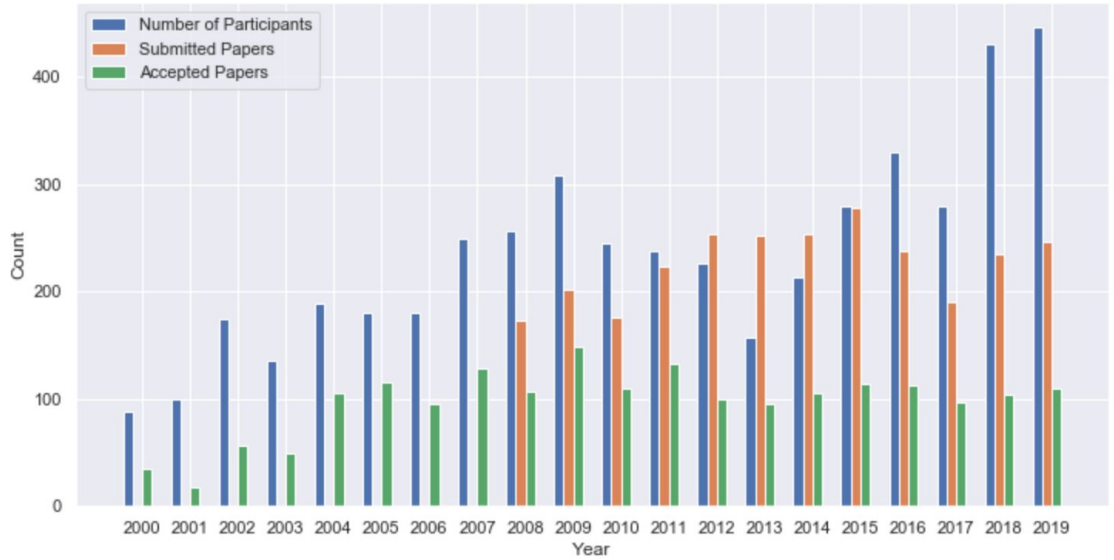
ISMIR

[International Society for Music Information Retrieval]

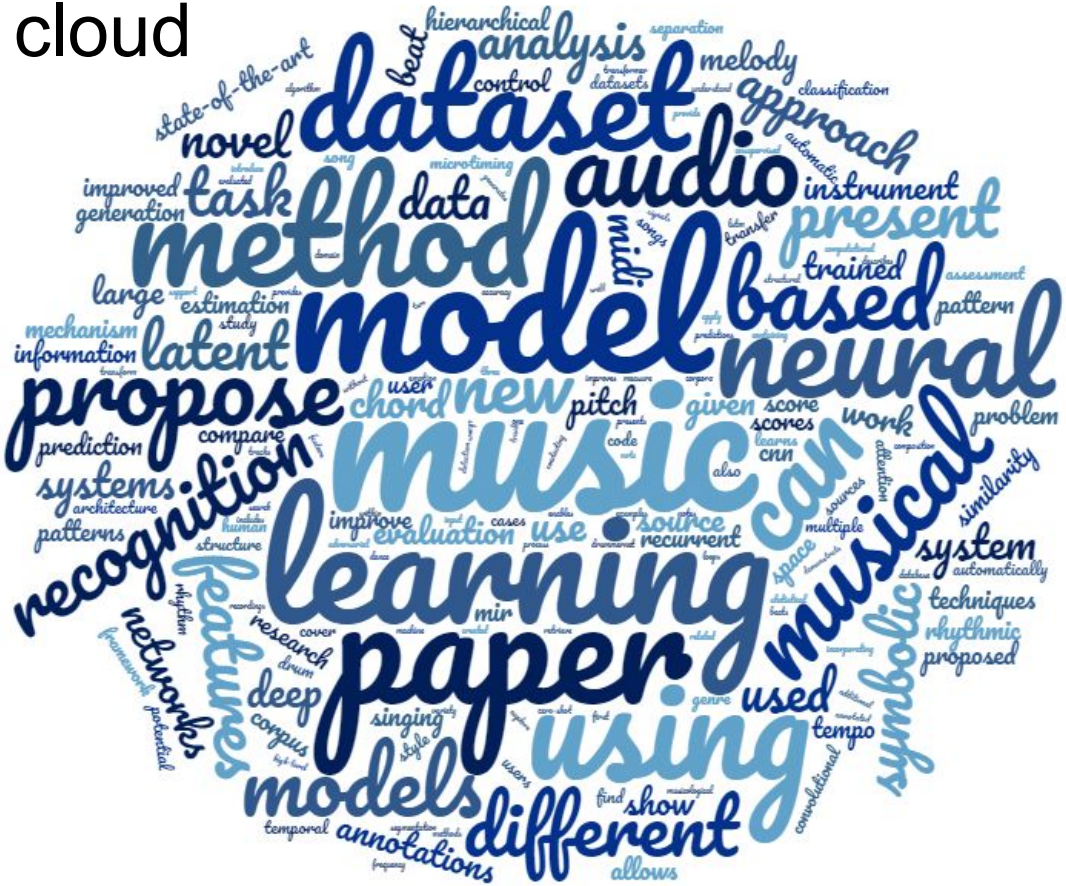
Annual, multidisciplinary conference, this year: 20th anniversary

Growing number of participants

~100 accepted papers



ISMIR tag cloud



Music Information Retrieval (MIR)

Common MIR tasks:

- Music transcription (melody, notes, instruments, music score)
- Rhythm analysis (beats, downbeats, tempo, rhythmic patterns, meter)
- Harmonic analysis (chords, key)
- Genre and mood identification
- Music similarity
- Music discovery and recommendation (playlist generation)
- Music generation
- ...

Learning Complex Basis Functions for Invariant Representations of Audio

Stefan Lattner, Monika Dörfler and Andreas Arzt

PDF

Typically: Fourier transformation for feature extraction

Complex-Autoencoder learns bases for orthogonal-transformation-invariant features

Magnitude-space: invariant; phase-space: variant

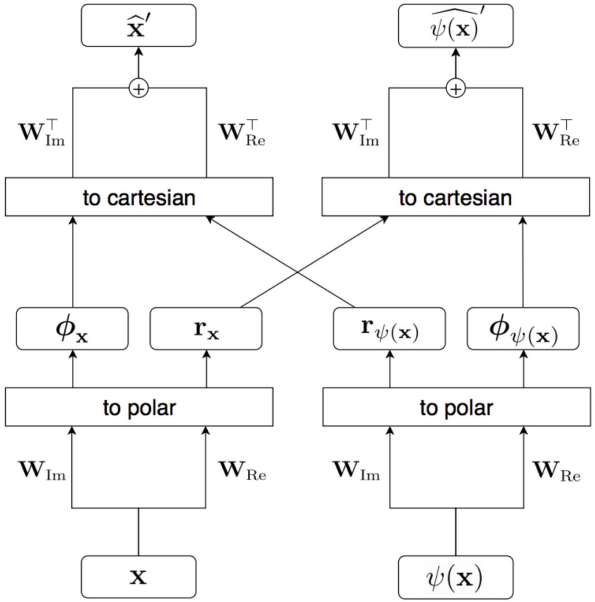
Learning Complex Basis Functions for Invariant Representations of Audio

Goal: learn imaginary / real submatrix of eigenvectors of orthogonal transformations, use as bases

How?

1. Map signal and transformed signal to current estimate
2. Express in polar coordinates
3. Reconstruct with swapped magnitude vectors
4. Check symmetric reconstruction error

Transformation: Difference in phase vectors



Towards Interpretable Polyphonic Transcription with Invertible Neural Networks

Rainer Kelz and Gerhard Widmer

PDF

Invertible networks unify discriminative and generative aspects in one function, i.e. they have one shared set of parameters

Allow direct inspection of what the discriminative model has learned

Predictions are interpretable, since it can be determined which inputs lead to which outputs

Towards Interpretable Polyphonic Transcription with Invertible Neural Networks

Invertible neural networks (INNs) produce models which:

- are able to derive semantic information from the input
- and are able to answer questions like *“is A a representative example for concept B?”*

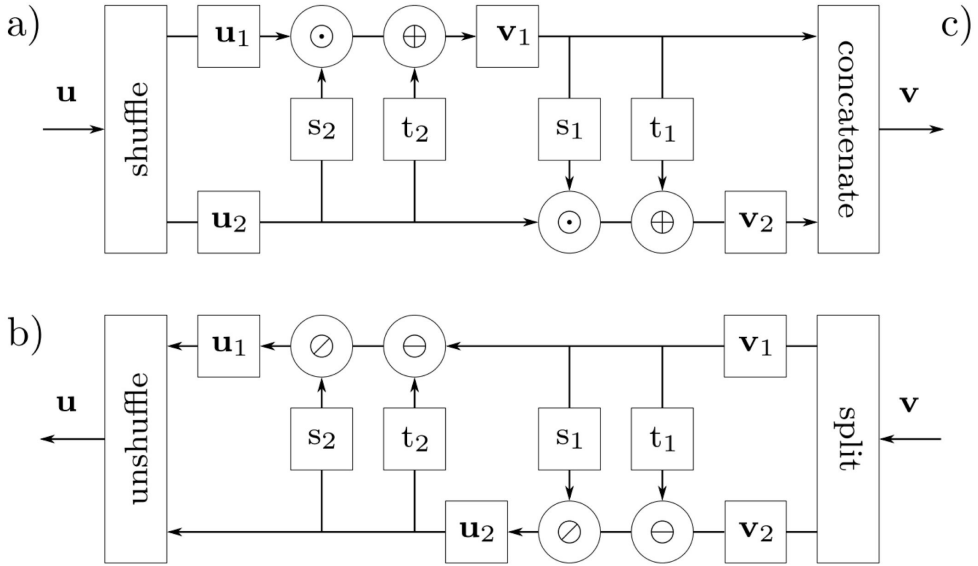
INNs are parameterised, nonlinear, bijective functions

Can be used to transform complex distributions into simple, factorised distributions

Trained similar to other NNs in a supervised fashion

Towards Interpretable Polyphonic Transcription with Invertible Neural Networks

Internal structure of invertible layer:



a) forward
b) backward
c) parameterisation of *s* and *t* transforms

$$\mathbf{v}_1 = \mathrm{cexp}(\mathrm{s}_2(\mathbf{u}_2)) \odot \mathbf{u}_1 \oplus \mathrm{t}_2(\mathbf{u}_2)$$
$$\mathbf{v}_2 = \mathrm{cexp}(\mathrm{s}_1(\mathbf{v}_1)) \odot \mathbf{u}_2 \oplus \mathrm{t}_1(\mathbf{v}_1)$$
$$\mathbf{u}_2 = (\mathbf{v}_2 \ominus \mathrm{t}_1(\mathbf{v}_1)) \oslash \mathrm{cexp}(\mathrm{s}_1(\mathbf{v}_1))$$
$$\mathbf{u}_1 = (\mathbf{v}_1 \ominus \mathrm{t}_2(\mathbf{u}_2)) \oslash \mathrm{cexp}(\mathrm{s}_2(\mathbf{u}_2))$$

Towards Interpretable Polyphonic Transcription with Invertible Neural Networks

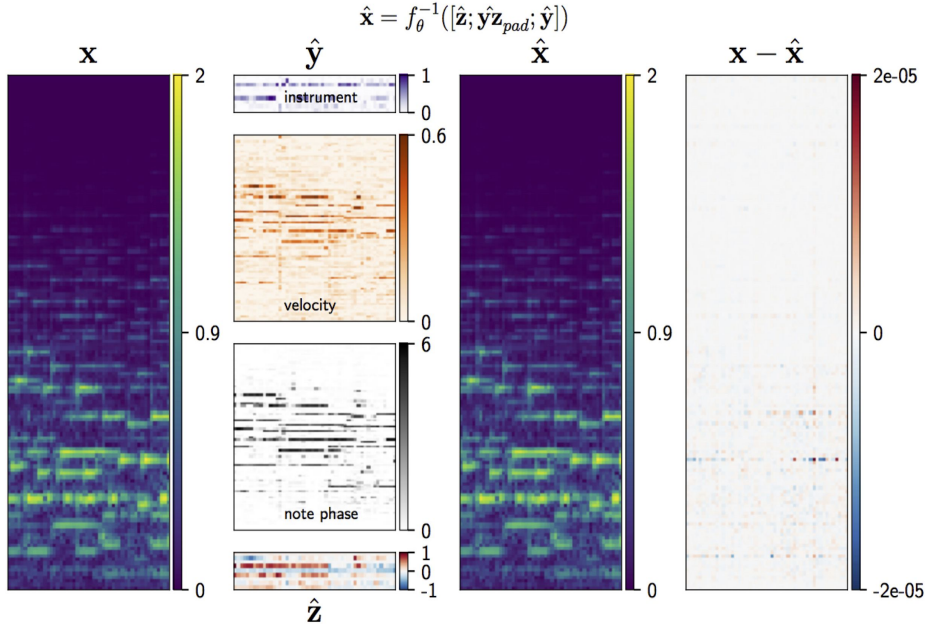
Transcription: multi-label problem

Input **X**: magnitude spectrogram

Output:

- **Y**: semantic part
  contains variables of interest
  note phase, velocity, and instrument
- **Z**: nuisance part
  contains all other irrelevant factors
  e.g. noise, reverberation, microphone

Input can be reconstructed given the output



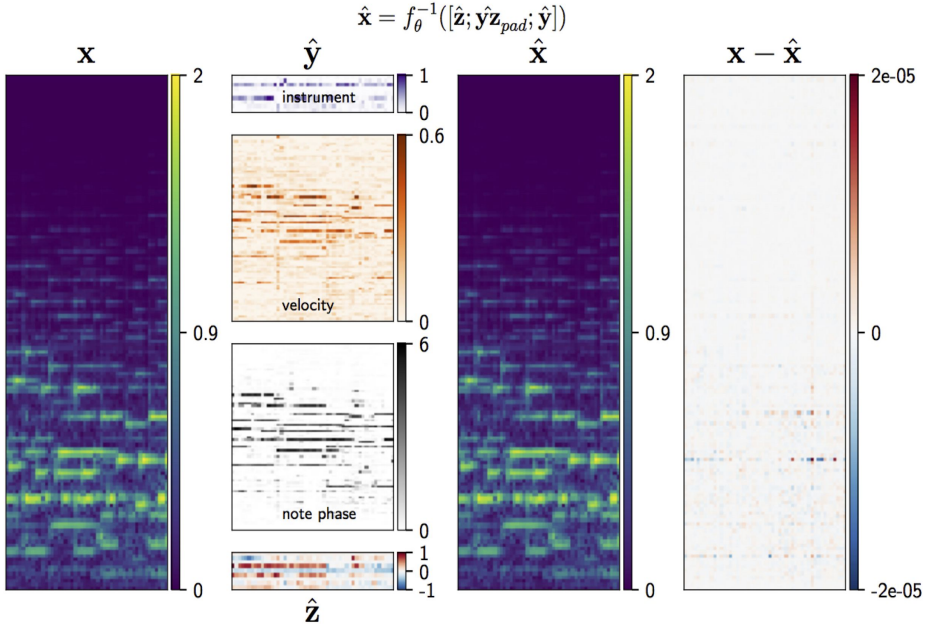
Towards Interpretable Polyphonic Transcription with Invertible Neural Networks

Interpretability

Every output has a directly interpretable correspondence in the input.

Assuming a perfect model, one would be able to:

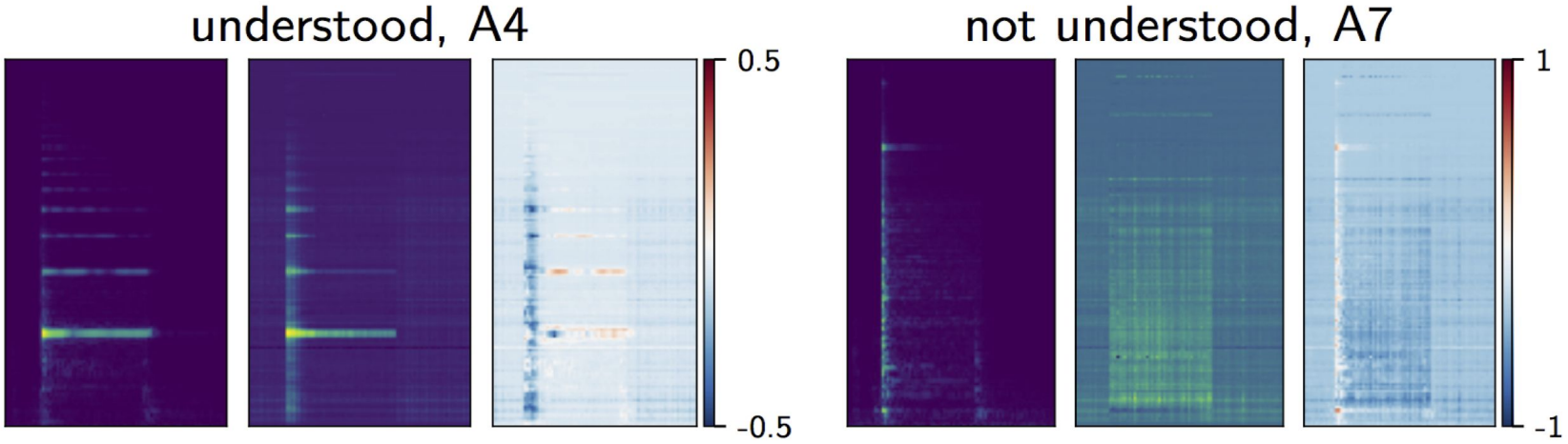
- add/delete notes,
- change velocity or instrument.



Towards Interpretable Polyphonic Transcription with Invertible Neural Networks

Concept Understanding

Since the model can be sampled to generate notes, it can be investigated which concepts are understood by the model (and which are not).



Adversarial Learning for Improved Onsets and Frames Music Transcription

Jong Wook Kim and Juan Bello

PDF

Automatic transcription of polyphonic music can be split into these main sub-problems:

1. Multi-pitch estimation (which pitches are sounding at time t),
2. Note tracking (models the evolution of the pitches to form notes).

Adversarial Learning for Improved Onsets and Frames Music Transcription

For multi-pitch estimation commonly CNNs and/or RNNs are used.

Most methods use an element-wise optimisation objective and do not account for inter-label dependencies (i.e. which notes/pitches do occur simultaneously).

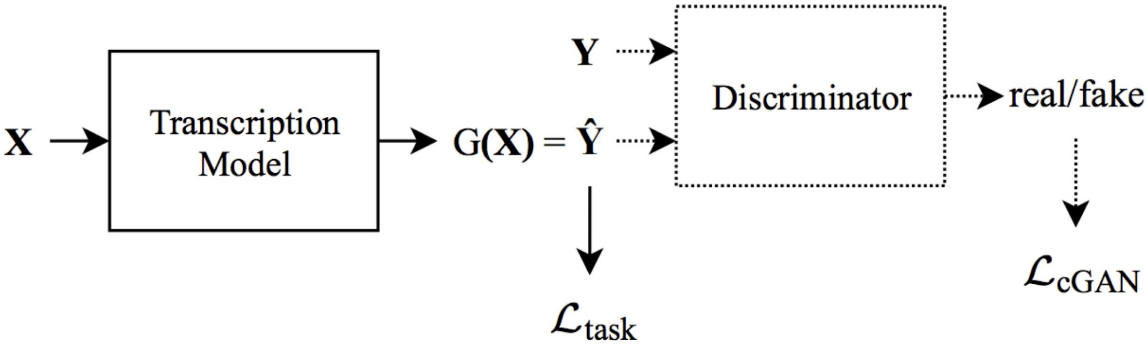
Proposed system applies an adversarial loss to reduce transcription errors.

Generative adversarial networks (GANs) consist of two components, the generator **G** and the discriminator **D**.

Adversarial Learning for Improved Onsets and Frames Music Transcription

GANs implement **G** and **D** as neural networks and train them in an adversarial manner, where:

1. the generator **G** learns to produce realistic samples,
2. and discriminator **D** learns to distinguish generated samples from real data.

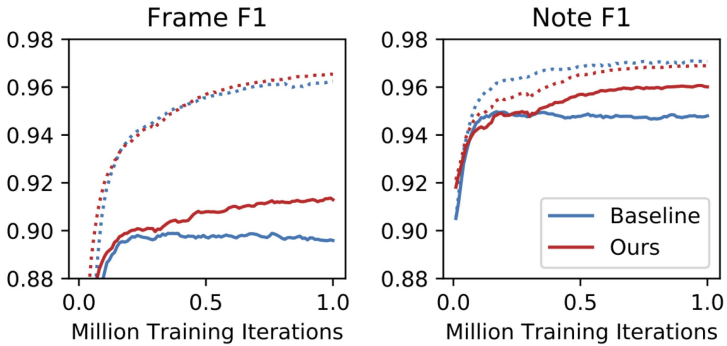


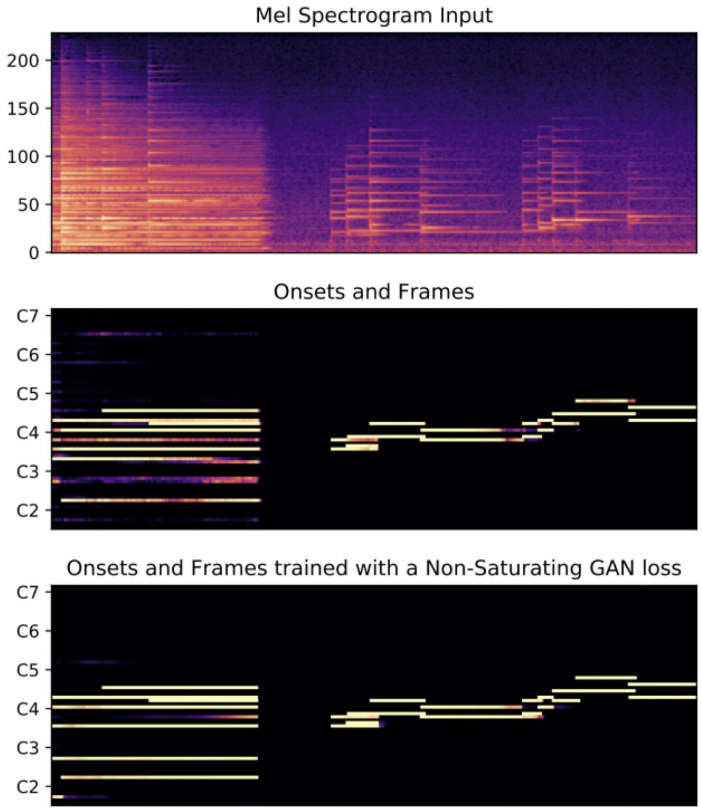
Adversarial Learning for Improved Onsets and Frames Music Transcription

Discriminator serves as a learned regulariser.

Improved note transcription: predicting more confident output.

Smaller generalisation gap:





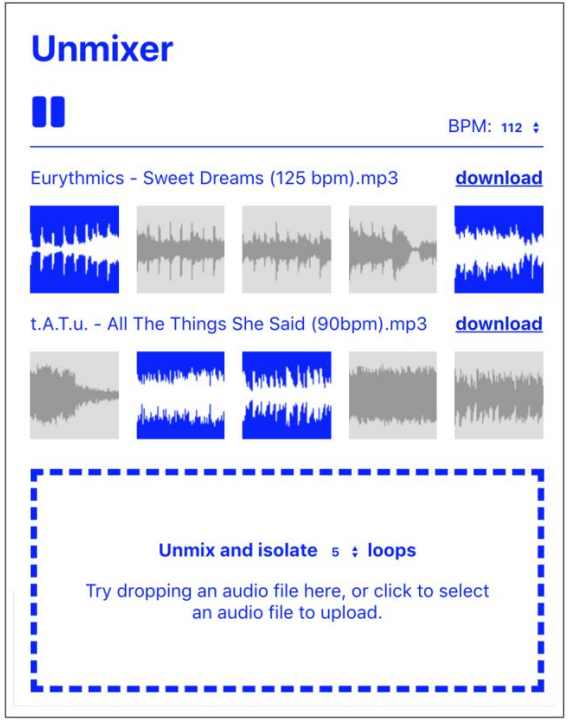
Unmixer: An Interface for Extracting and Remixing Loops

Jordan Smith, Yuta Kawasaki and Masataka Goto

PDF

Unmixer Interface

Isolates a defined number of loops

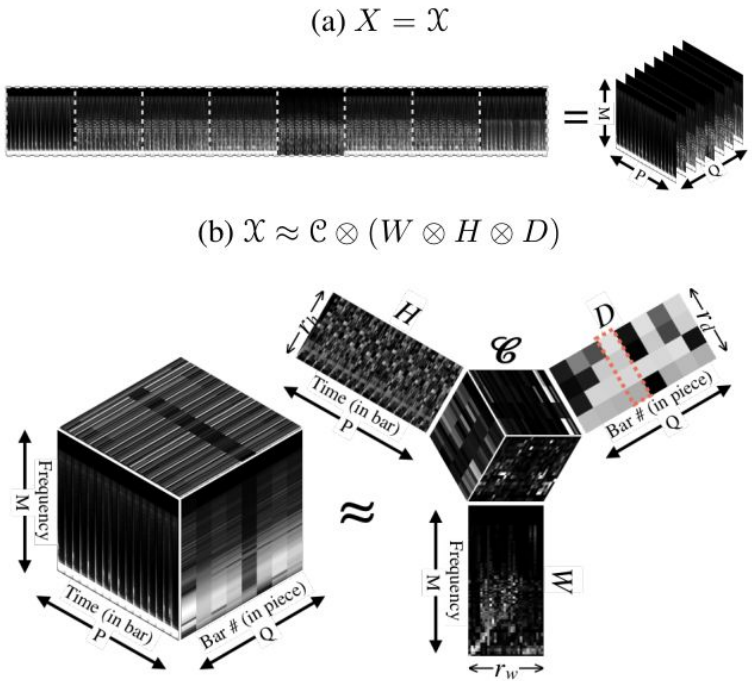


Unmixer: An Interface for Extracting and Remixing Loops

How are isolated loops found?

1. Compute spectrogram of audio
2. Divide into downbeat-sized windows
3. Stack spectrogram windows per bar (a)
4. Compute non-negative Tucker decomposition (b)

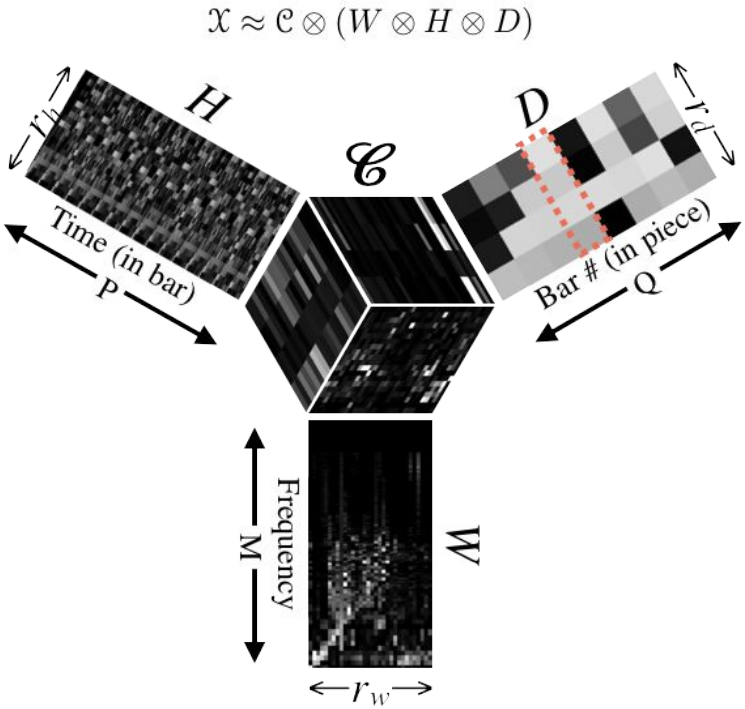
Results in: spectral templates **W**, time-activation templates **H** and loop-activation templates **D**



Unmixer: An Interface for Extracting and Remixing Loops

After Tucker decomposition:

5. Adapt core tensor with sparsity constraint in 3rd dimension
6. Compute contribution of k-th loop and unfold into real-valued spectrum
7. Reconstruction of signal
8. Select most prominent bar for each loop



ISMIR conference dinner impressions

Photo credit:
Christof Weiss