Multi-voice Polyphonic Music Transcription Using Eigeninstruments

Graham Grindlay and Daniel P. W. Ellis

LabROSA, Dept. of Electrical Engineering, Columbia University

{grindlay,dpwe}@ee.columbia.edu

Overview

- System for transcribing **multi-instrument**, polyphonic musical recordings
- Implicitly handles source (instrument) separation
- Based on novel semi-supervised NMF variant called *Subspace NMF* (SsNMF)
- SsNMF incorporates prior knowledge by imposing constraints derived from training data

Non-negative Matrix Factorization for Music Transcription

- Non-negative matrix factorization (NMF) solves $V \approx WH$ [1]
- One possible error function function (generalized KL-divergence):

$$D(V||WH) = \sum_{i=1}^{f}\sum_{j=1}^{t}\left(V_{ij}\log\frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}\right)$$

- Fast multiplicative updates exist to solve for W and H:

$$W_{ik} \leftarrow W_{ik}\frac{\sum_j H_{kj}\frac{V_{ij}}{(WH_{ij})}}{\sum_j H_{kj}} \qquad H_{kj} \leftarrow H_{kj}\frac{\sum_i W_{ik}\frac{V_{ij}}{(WH_{ij})}}{\sum_i W_{ik}}$$

- Smaragdis and Brown showed how NMF can be used for piano music transcription [2]
- V is the f-by-t magnitude STFT of the audio
- W contains note spectra in its columns and represents a source model
- H contains note activations in its rows and gives the transcription
- Rank of decomposition corresponds to number of pitches p
- W unknown *a priori* $\rightarrow$ unsupervised transcription
- W known *a priori* $\rightarrow$ supervised transcription

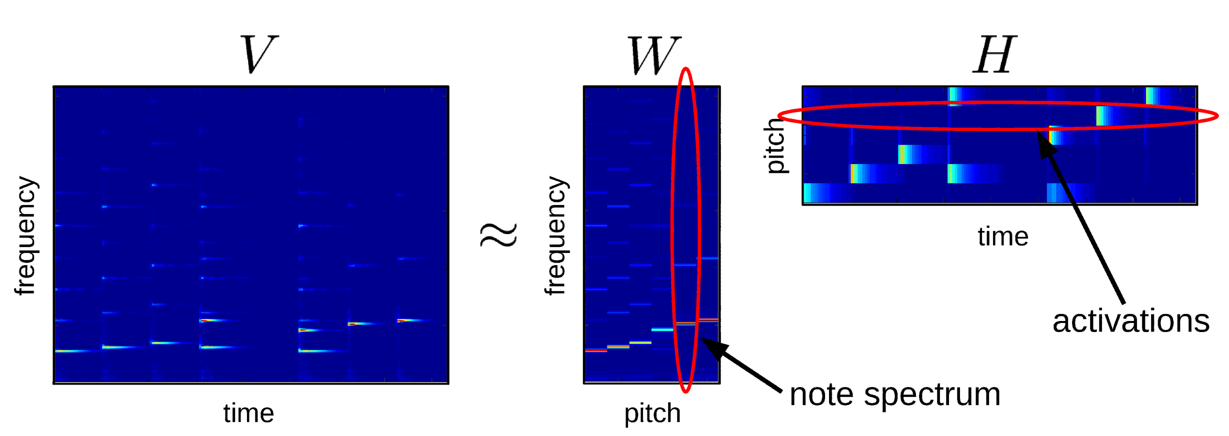

Figure 1: Using NMF to transcribe a piano note sequence (pitches have been manually sorted)

- Can extend to mixtures of n sources (instruments) by interpreting W and H in block-form:

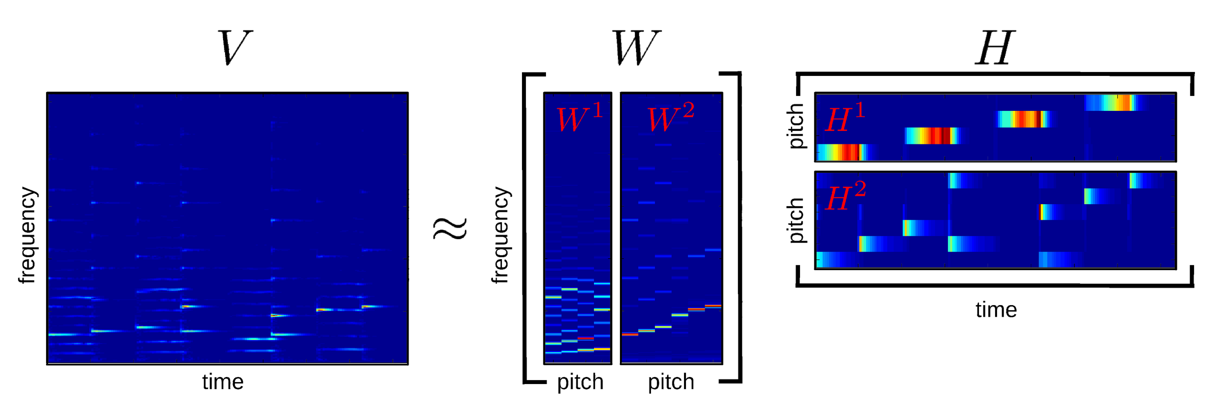

Figure 2: Using NMF to transcribe a mixture of piano and cello

- Not clear how to assign columns of W to the submatrices W^i in the unsupervised case!

Subspace NMF

- **Idea**: Constrain solution of each W^i to lie in a linear subspace derived from training data
- Reminiscent of "eigenvoice" technique used in speech recognition [3, 4]

Training

- Given set of m instrument models $\mathcal{M}$, each with p pitches and f frequency bins
- Vectorize models and and combine into a model matrix $\Theta = [vec(\mathcal{M}^1)\ vec(\mathcal{M}^2)\ldots vec(\mathcal{M}^m)]$
- Decompose model matrix using rank-r NMF: $\Theta \approx \Omega C$
- Unvectorize model basis vectors: $\mathcal{W}^i = vec^{-1}(\Omega_i)$
- Each $\mathcal{W}^i$ represents an "eigeninstrument" (f-by-p matrix)

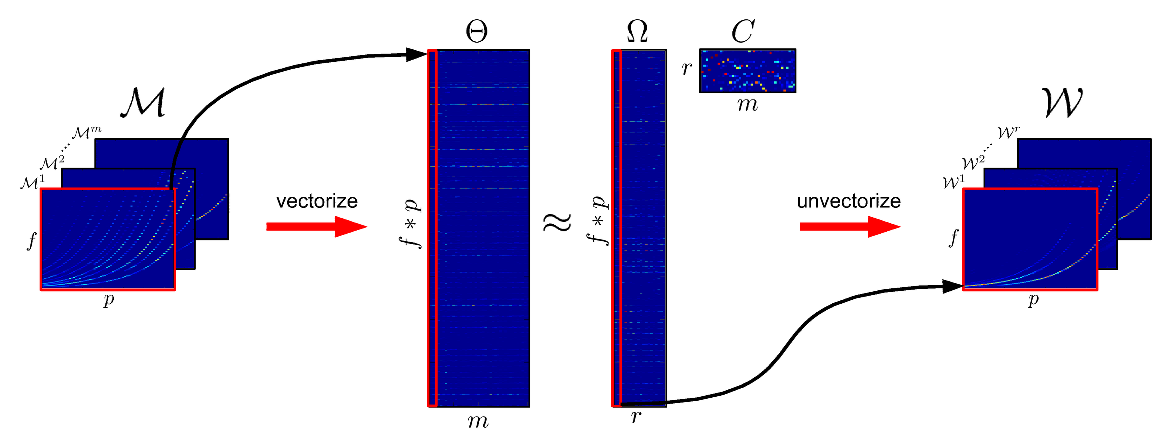

Figure 3: Process of deriving "eigeninstruments" from a set of training instrument models

The Model

- Use eigeninstrument basis to represent mixture of n unknown instruments V as:

$$V \approx \sum_{s=1}^{n} W^s H^s = \sum_{s=1}^{n}\sum_{a=1}^{r}\mathcal{W}^a B_{as} H^s$$

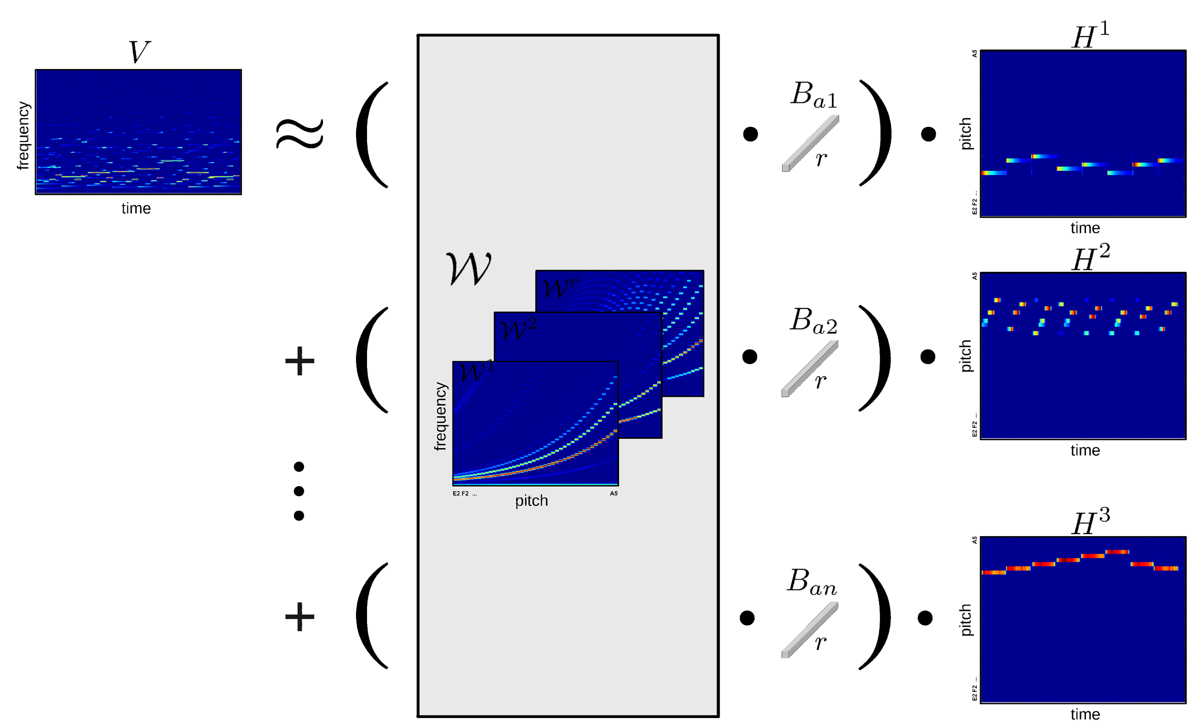

Figure 4: Illustration of the Subspace NMF decomposition of a spectrogram

Transcription

1. Update each H^s by combining into big H and using NMF update
2. Update for B is as follows:

$$B_{as} \leftarrow B_{as}\frac{\sum_{i=1}^{f}\sum_{j=1}^{t} V_{ij}\frac{\sum_{k=1}^{p}\mathcal{W}_{ik}^{a}H_{kj}^{s}}{\sum_{s=1}^{n}\sum_{b=1}^{r}\sum_{k=1}^{p}\mathcal{W}_{ik}^{b}B_{bs}H_{kj}^{s}}}{\sum_{i=1}^{f}\sum_{j=1}^{t}\sum_{k=1}^{p}\mathcal{W}_{ik}^{a}H_{kj}^{s}}$$

3. Solve for each W^s using B
4. Iterate until convergence
5. Post-process H^s using median filtering and thresholding to get pianoroll representation

Experiments

- Experiments conducted with both synthesized (MIDI) and audio recordings
- MIDI-derived instrument models used as training data
- Frame-level metrics: total error, substitutions, missed notes, false alarms, accuracy

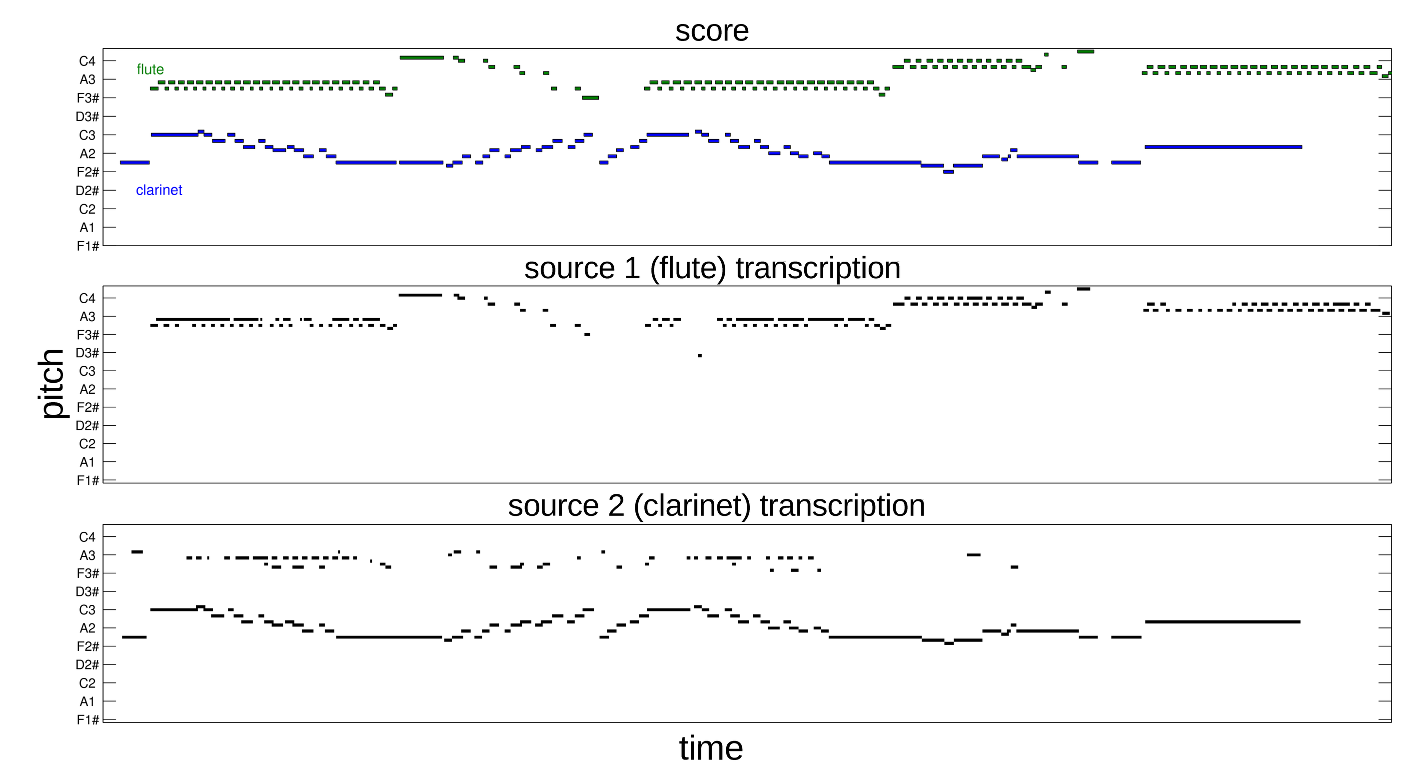

Figure 5: Transcription results of Beethoven string quartet recording (two sources)

	Acc	E_{tot}	E_{sub}	E_{miss}	E_{fa}
Recorded Audio (flute & clarinet)	0.65	0.43	0.04	0.11	0.28
Synthesized Audio (bass & piano)	0.69	0.32	0.07	0.11	0.13
Synthesized Audio (flute & violin)	0.72	0.31	0.03	0.18	0.11

Table 1: Experimental results (averaged across sources) of three mixtures, each with two sources

Discussion

- SsNMF provides a framework for transcribing multi-instrument, polyphonic recordings
- Adaptive source modeling has distinct advantages over a purely supervised approach
- Current work involves extending the static spectrum note model to handle dynamic spectra

References

[1] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2001, pp. 556–562.

[2] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[3] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker identification in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, November 2000.

[4] R. Weiss and D. Ellis, "Monaural speech separation using source-adapted models," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 114–117.