A Probabilistic Subspace Model for Multi-Instrument Polyphonic Transcription

Graham Grindlay and Daniel P. W. Ellis

LabROSA, Dept. of Electrical Engineering, Columbia University

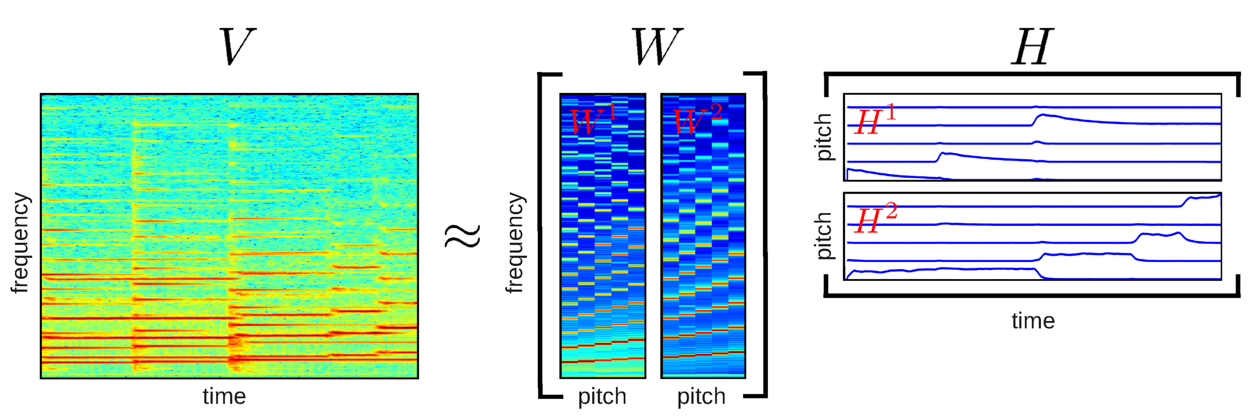
{grindlay,dpwe}@ee.columbia.edu

Overview

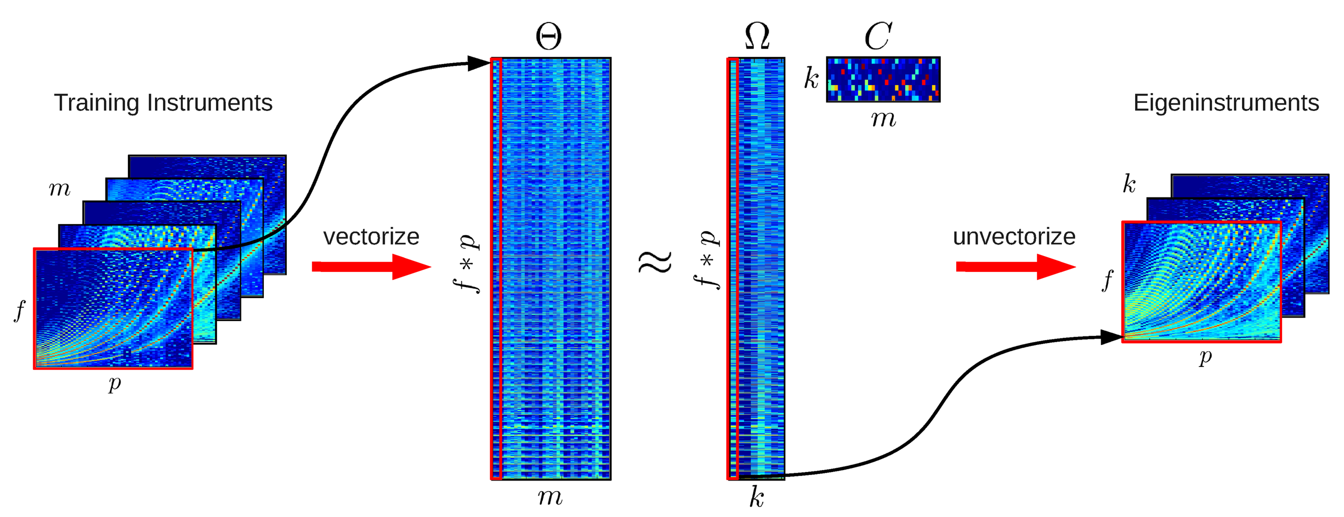
- System for transcribing polyphonic music
  - Works with single channel audio
  - Handles **multiple instruments**
- Probabilistic extension of Subspace NMF [1]
- Prior knowledge included via parametric instrument models
- Constraints derived from example instrument models
- Can initialize model with correct instrument types
- Sparsity heuristic improves performance

Subspace NMF for Music Transcription

- Non-negative matrix factorization (NMF) [3] solves $V \approx WH$
- V is the f-by-t magnitude STFT of the audio
- W contains note spectra in its columns and represents the source model(s)
- H contains note activations in its rows and gives the transcription
- Rank of decomposition corresponds to number of distinct pitches
- Can handle multiple instruments by interpreting W and H in block-form (i.e. $V \approx \sum_s W^s H^s$)



- **Problem**: blind search for instrument models W is an ill-posed problem
- Subspace NMF [1] constrains each W^s to lie in a subspace derived from training data



- Given set of m instrument models for training, each with p pitches and f frequency bins
- Decompose matrix of training instrument parameters using rank-k NMF: $\Theta \approx \Omega C$
- Instrument sources can be represented as linear combinations of the eigeninstruments basis:

$$V \approx \sum_s vec^{-1}(\Theta \mathbf{b}_s) H^s$$

Probabilistic Eigeninstrument Transcription

- NMF has probabilistic interpretation as latent variable model [2, 4]

$$V = P(f,t) \approx P(t) \sum_z P(f|z)P(z|t)$$

- *Probabilistic Eigeninstrument Transcription* (PET) generalizes Subspace NMF in a similar way:

$$V = P(f,t) \approx P(t) \sum_{s,p,k} \hat{P}(f|p,k)P(k|s)P(s|p,t)P(p|t)$$

$\hat{P}(f\|p,k)$	-	probabilistic eigeninstruments model
$P(k\|s)$	-	source-specific distribution over eigeninstruments
$P(s\|p,t)$	-	probability of instrument source s playing pitch p at time t
$P(p\|t)$	-	probability of pitch p at time t

PET Algorithm

1. Calculate probabilistic eigeninstruments $\hat{P}(f|p,k)$ from training data
2. Solve model parameters for a given test mixture V using EM
3. Compute joint pitch-time distribution for each source:

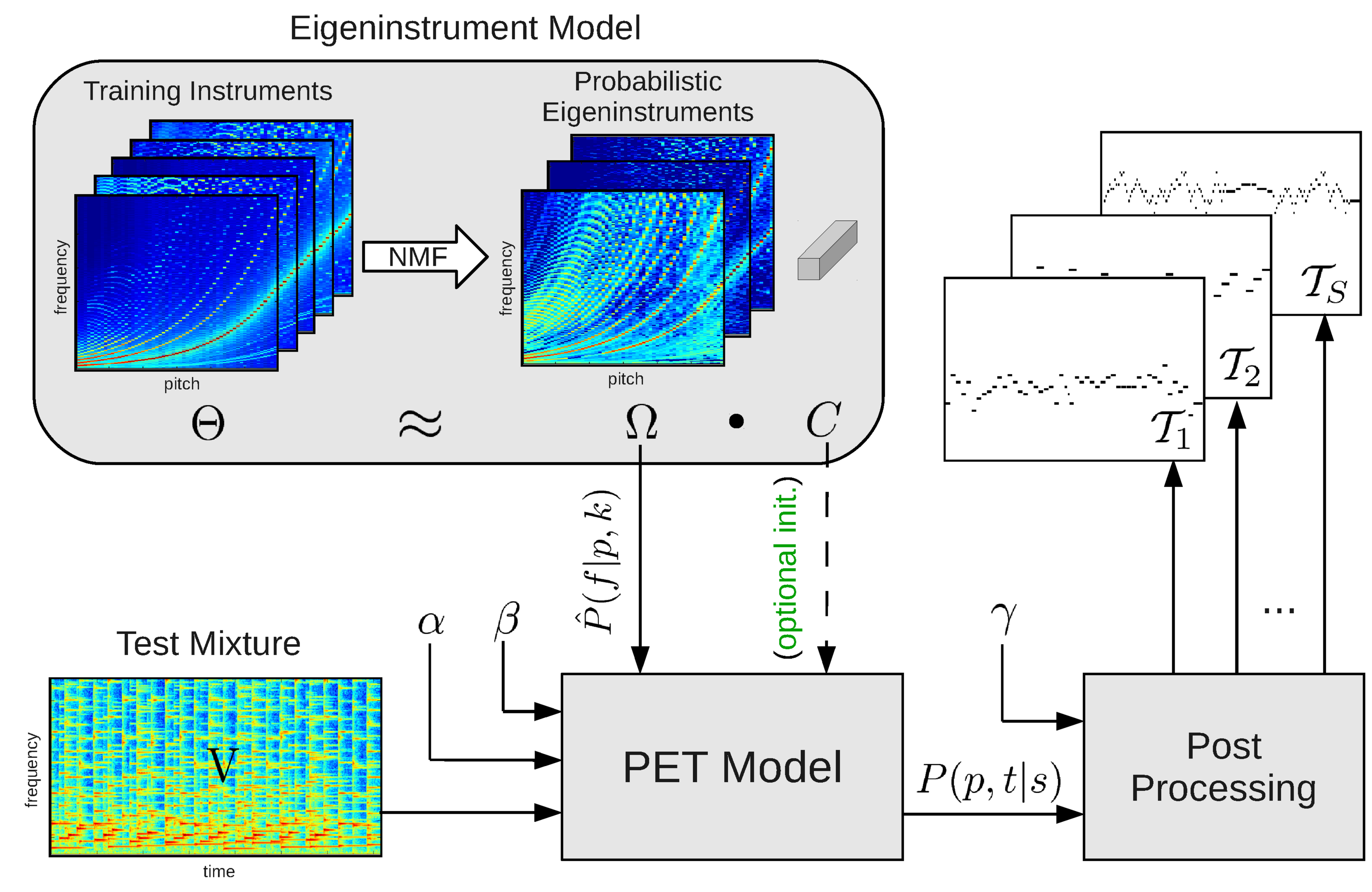
$$P(p,t|s) = \frac{P(s|p,t)P(p|t)P(t)}{\sum_{p,t} P(s|p,t)P(p|t)P(t)}$$

4. Post-process $P(p,t|s)$ into binary piano roll $\mathcal{T}_s$ using threshold γ

- Can encourage **sparsity** in latent variable distributions using exponentiation heuristic in M-step:

$$\boxed{P(s|p,t) = \frac{\left[\sum_{f,k} P(s,p,k|f,t)V_{f,t}\right]^\alpha}{\sum_s \left[\sum_{f,k} P(s,p,k|f,t)V_{f,t}\right]^\alpha}} \qquad \boxed{P(p|t) = \frac{\left[\sum_{f,k,s} P(s,p,k|f,t)V_{f,t}\right]^\beta}{\sum_p \left[\sum_{f,k,s} P(s,p,k|f,t)V_{f,t}\right]^\beta}}$$

- Semi-supervised variant: initialize $P(k|s)$ with eigeninstrument weights from **similar** instrument types



Experiments

- Tested two-instrument mixtures from two synthesized data sets and one recorded data set
- Eigeninstruments generated from set of 33 training instruments (wide variety of types)
- Threshold γ was derived empirically to maximize F-measure across tracks
- Evaluated basic PET model, PET with sparsity on $P(s|p,t)$, and PET with sparsity on $P(p|t)$
- Compared to oracle/fixed-model (synthesized/recorded data) PET system and to NMF with generic W model

	Bach (synth)	Woodwind (synth)	Woodwind (recorded)
PET	0.51	0.53	0.50
$PET_{\alpha=2}$	0.57	0.62	0.52
$PET_{\beta=2}$	0.52	0.51	0.56
PET_{init}	0.56	0.68	0.59
$PET_{oracle/fixed}$	0.87	0.84	0.56
NMF	0.41	0.40	0.31

Table 1: Frame-level F-measures across three data sets for PET variants as well as basic NMF

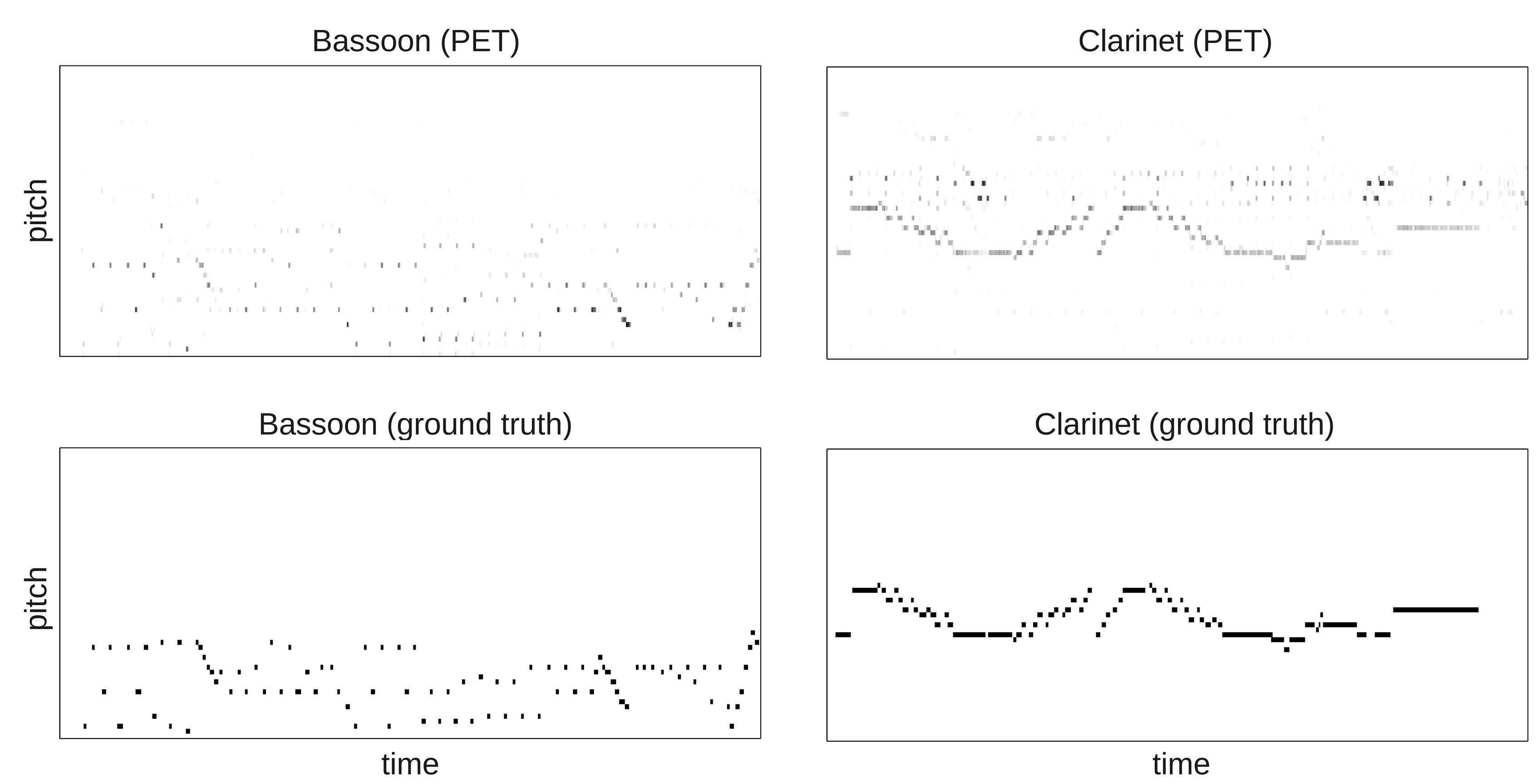


Figure 1: Example output distributions $P(p,t|s)$ and ground truth for a recorded bassoon-clarinet mixture

Discussion

- Sparsity heuristic is helpful in most situations, although different data sets benefit in different ways
- Initializing model with approximately correct parameters can improve accuracy
- PET framework shows significant performance advantage over basic NMF algorithm

References

[1] G. Grindlay and D.P.W. Ellis. Multi-voice polyphonic music transcription using eigeninstruments. In *WASPAA*, 2009.

[2] T. Hofmann. Probabilistic latent semantic analysis. In *Uncertainty in AI*, 1999.

[3] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, 2001.

[4] M. Shashanka, B. Raj, and P. Smaragdis. Probabilistic latent variable models as non-negative factorizations. *Computational Intelligence and Neuroscience*, 2008.