A Probabilistic Subspace Model for Multi-Instrument Polyphonic Transcription

Graham Grindlay and Daniel P. W. Ellis
LabROSA, Dept. of Electrical Engineering, Columbia University
{grindlay,dpe}@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 \( f \times t \) magnitude STFT of the audio
- \( W \) contains note spectra in its columns and represents the source models
- \( H \) 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 \( W \) in block-form (i.e. \( V \approx \sum W^c H^c \))

Problem

- Blind search for instrument models \( W \) is an ill-posed problem
- Subspace NMF [1] constrains each \( H^c \) 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 \( \Theta = W H \)
- Instrument sources can be represented as linear combinations of the eigeninstruments basic
  \( V \approx \sum \text{vec}^{-1}(W) \Theta k \)

Probabilistic Eigeninstrument Transcription

- NMF has probabilistic interpretation as latent variable model [2, 4]
  \[ V = P(f,t) \approx P(t) \sum_{s,p,k} P(s|p)P(t|p)P(p|t) \]
- Probabilistic Eigeninstrument Transcription (PET) generalizes Subspace NMF in a similar way:
  \[ V = P(f,t) \approx P(t) \sum_{s,p,k} P(s|p,k)P(k|s)P(k|t)P(t|p) \]

PET Algorithm

1. Calculate probabilistic eigeninstruments \( P(s|p,k) \) from training data
2. Solve model parameters for a given test mixture \( V \) using EM
3. Compute joint pitch-time-distribution for each source:
   \[ P(p|t) = \frac{P(p|t)}{\sum_{s,p,k} P(s|p,k)P(k|t)P(t|p)} \]
4. Post-process \( P(p|t) \) into binary piano roll \( T_S \) using threshold \( \gamma \)
- Can encourage sparsity in latent variable distributions using exponentiation heuristic in M-step:
  \[ P(p|t) = \left( \frac{\sum_{s,p,k} P(s,p,k|f,TV_f)}{\sum_{s,t} P(s,p,k|f,TV_f)} \right)^\beta \]
- Semi-supervised variant: initialize \( P(s|p,k) \) with eigeninstrument weights from similar instrument types

Experimental Results

- 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 \( \gamma \) was derived empirically to maximize F-measure across tracks
- Evaluated basic PET model, PET with sparsity on \( P(p|t) \), and PET with sparsity on \( P(s|p,k) \)
- Compared to oracle/fixed-model (synthesized/recorded data) PET system and to NMF with generic W model

<table>
<thead>
<tr>
<th></th>
<th>Back (sympt)</th>
<th>WoBraid (sympt)</th>
<th>WoBraid (rec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PET</td>
<td>0.51</td>
<td>0.53</td>
<td>0.56</td>
</tr>
<tr>
<td>PETm=2</td>
<td>0.77</td>
<td>0.62</td>
<td>0.52</td>
</tr>
<tr>
<td>PETm=4</td>
<td>0.55</td>
<td>0.51</td>
<td>0.56</td>
</tr>
<tr>
<td>PETm=2/fixed</td>
<td>0.76</td>
<td>0.68</td>
<td>0.59</td>
</tr>
<tr>
<td>NMF</td>
<td>0.41</td>
<td>0.40</td>
<td>0.31</td>
</tr>
</tbody>
</table>

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

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