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):
  \[
  D(V||WH) = \sum_{i=1}^{n} \sum_{j=1}^{m} (V_{ij} \log \frac{V_{ij}}{W_iH_j})
  \]
- Fast multiplicative updates exist to solve for \( W \) and \( H \)
  \[
  W_{ik} = \frac{H_k}{\sum_{j=1}^{m} H_{kj}} \]
  \[
  H_{kj} = \frac{\sum_{i=1}^{n} V_{ij} W_{i1}}{\sum_{i=1}^{n} \sum_{j=1}^{m} W_{ij} H_{kj}}
  \]
- Sminaghi and Ksoum 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](image1.png)

\( V \)  \( H \)  \( W \)

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

![Figure 2](image2.png)

Using NMF to transcribe a mixture of piano and cello

![Figure 3](image3.png)

Process of deriving "eigeninstruments" from a set of training instrument models

![Figure 4](image4.png)

Illustration of the Subspace NMF decomposition of a spectrogram

Transcription

1. Update each \( H \) by combining into big \( H \) and using NMF update
2. Update for \( H \) as follows:
   \[
   H_{kj} = \frac{\sum_{i=1}^{n} V_{ij} W_{i1}}{\sum_{i=1}^{n} \sum_{j=1}^{m} W_{ij} H_{kj}}
   \]
3. Solve for each \( W^* \) using \( H \)
4. Iterate until convergence
5. Post-process \( H \) using median filtering and thresholding to get pianoroll representation

Subspace NMF

- Idea: Constrain solution of each \( W^* \) 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 \( M_i \) each with \( p \) pitches and \( f \) frequency bins
- Vectorize models and and combine into a model matrix \( \Phi = \text{vec}(M_1^T \cdots \text{vec}(M_m^T)) \)
- Decompose model matrix using ranks \( NMF \) \( \Phi \approx BC \)
- Uncorrelated model basis vectors: \( W^* = \text{vec}^{-1}(B) \)
- Each \( W^* \) represents an "eigeninstrument" (\( f \)-by-\( p \) matrix)

![Figure 5](image5.png)

Transcription results of Rossetti string quartet recording (two sources)

The Model

- Use eigeninstrument basis to represent mixture of \( n \) unknown instruments \( V \) as:
  \[
  V \approx \sum_{s=1}^{n} W^*_s B_s H^*_s
  \]

![Table 1](image6.png)

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