Factorial Scaled Hidden Markov Model for Polyphonic Audio Representation and Source Separation

Alexey Ozerov (1), Cédric Févotte (2) and Maurice Charbit (3)

19 October 2009

(1) INRIA / IRISA – Rennes, France (this work was done while in Telecom ParisTech),
(2) CNRS LTCI, Telecom ParisTech, France
(3) Institut Telecom, Telecom ParisTech, France
Introduction

- We consider two state-of-the-art models for polyphonic audio representation:
  - Itakura-Saito nonnegative matrix factorization (IS-NMF)
  - Gaussian scaled mixture model (GSMM)

- We combine both models into a hybrid model:
  - Factorial scaled hidden Markov model (FS-HMM)

- We apply FS-HMM to single-channel speech / music separation
Outline

- State-of-the-art models
- Motivation
- Factorial scaled hidden Markov model
- Inference algorithms
- Application to single-channel speech / music separation
- Conclusion
State-of-the-art models

- Family of models considered: short time spectra are modeled as Gaussians with zero means and structured variances

$$X_{fn} \sim N_c(0, \nu_{fn})$$

$$V = \left(\nu_{fn}\right)_{fn}$$

- Attractive properties:
  - Generative model (for source separation)
  - Model is linear in STFT domain
    - Easy inference
    - Easy signal reconstruction
State-of-the-art models

- Itakura-Saito nonnegative matrix factorization (IS-NMF) [Benaroya et al. 2003, Févotte et al. 2009]
  - Suitable for polyphonic signals

- Gaussian scaled mixture model (GSMM) [Benaroya et al. 2006]
  - Suitable for monophonic signals
State-of-the-art models

- IS-NMF

\[ p_{X_f}(x) = N_c \left( x; 0, \sum_{k=1}^{K} w_{fk} h_{kn} \right) \]

- GSMM

\[ p_{X_f}(x) = \sum_{j=1}^{J} c_j N_c \left( x; 0, w_j, fh_{j,n} \right) \]

- Summation of variances

- Summation of probability density functions (pdfs)
Outline

- State-of-the-art models
- Motivation
- Factorial scaled hidden Markov model
- Inference algorithms
- Application to single-channel speech / music separation
- Conclusion
Motivation

- We would like to marry these two models for the following reasons
  - Using suitable models for the corresponding sources (monophonic or polyphonic)
  - Introducing discrete states into IS-NMF
    - Facilitates the modeling of temporal dependencies
    - Leads to joint (or integrated) approaches for source separation and information retrieval (e.g., music transcription [Bertin et al 2007])
Outline

- State-of-the-art models
- Motivation
- Factorial scaled hidden Markov model
- Inference algorithms
- Application to single-channel speech / music separation
- Conclusion
Factorial scaled Hidden Markov model

- Mother (IS-NMF)
- Father (GSMM)
Factorial scaled Hidden Markov model

- Baby (Factorial scaled GMM or HMM)
Factorial scaled Hidden Markov model

Mother (IS-NMF)

\[ p_{X_{fn}}(x) = N_c(x; 0, \sum_{k=1}^{K} w_{fk} h_{kn}) \]

Father (GSMM)

\[ p_{X_{fn}}(x) = \sum_{i=1}^{J} \alpha_i N_c(x; 0, w_{i,f} h_{i,n}) \]

Baby 1 (FS-GMM)

\[ p_{X_{fn}}(x) = \left[ \sum_{i_1=1}^{J_1} \alpha_{i} N_c(x; 0, w_{i_1,f1} h_{i_1,1n}) \right] \otimes \ldots \otimes \left[ \sum_{i_K=1}^{J_K} \alpha_{i} N_c(x; 0, w_{i_K,fK} h_{i_K,Kn}) \right] \]

Baby 2 (FS-HMM)

We add temporal dependencies between states: first order Markov chain
Factorial scaled Hidden Markov model

Mother (IS-NMF)

\[ c_{k,n} \sim \mathcal{N}_c(0, h_{k,n} \text{diag}(w_k)) \]

Father (GSMM)

\[ x_n \sim \mathcal{N}_c(0, h_{i,n} \text{diag}(w_i)) \]

given \( I_n = i \)

\[ x_n \in \mathbb{C}^F \]

\[ c_{k,n} \in \mathbb{C}^F \]
Factorial scaled Hidden Markov model

Baby 1 (FS-GMM)

\[
c_{k,n} \sim \mathcal{N}_c(0, h_{i,kn} \text{diag}(w_{i,k})) \quad \text{given} \quad I_{k,n} = i
\]
Outline

- State-of-the-art models
- Motivation
- Factorial scaled hidden Markov model
- Inference algorithms
- Application to single-channel speech / music separation
- Conclusion
Inference algorithms

- Two Generalized EM (GEM) algorithms
- EM algorithm
  - Full complete data set:
    \( \mathcal{Y} = \{C, I\} \)

- EM-MU algorithm
  - Reduced complete data set:
    \( \mathcal{Z} = \{X, I\} \)
Inference algorithms

- Convergence speeds

![Graph showing convergence speeds of EM and EM-MU algorithms with log-likelihood values declining over iterations.]
Outline

- State-of-the-art models
- Motivation
- Factorial scaled hidden Markov model
- Inference algorithms
- Application to single-channel speech / music separation
- Conclusion
Application to single channel speech / music separation

\[
x_n = s_n + m_n, \quad s_n = \sum_{k=1}^{K_s} c_{k,n}, \quad m_n = \sum_{k=K_s+1}^{K_s+K_m} c_{k,n},
\]

- Tested model (FS-HMM) configurations:
  - **Mono. speech / Mono. music** (S-HMM / S-HMM):
    - \( K_s = K_m = 1 \) (\( K = 2 \)), \( J_1 = 16 \), \( J_2 = 8 \)
  - **Mono. speech / Poly. music** (S-HMM / IS-NMF):
    - \( K_s = 1 \), \( K_m = 8 \) (\( K = 9 \)), \( J_1 = 16 \), \( J_k = 1 \) (\( k > 1 \))
  - **Poly. speech / Poly. music** (IS-NMF / IS-NMF):
    - \( K_s = 16 \), \( K_m = 8 \) (\( K = 24 \)), \( J_k = 1 \) (\( k = 1, \ldots, K \))
Application to single channel speech / music separation

- **Data**
  - speech
  - music
  - mix

- **Procedure:**
  - Learn a speech model (from some training data)
  - Clamp the speech model spectral patterns and estimate all the other parameters (from the mix)
  - Reconstruct sources (via MMSE estimation)

Data

- TIMIT (10 male speakers / 10 female speakers)
- 10 music 1 min. experts

Procedure:
Application to single channel speech / music separation

- LEGO-like (e.g., modular framework)
Application to single channel speech / music separation

- Female speech model spectral patterns

**FS-HMM** (monophonic)  
**IS-NMF** (polyphonic)
Application to single channel speech / music separation

- Numerical results

<table>
<thead>
<tr>
<th>Speech model</th>
<th>S-HMM</th>
<th>S-HMM (in braces)</th>
<th>NMF</th>
<th>NMF (in braces)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Music model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male (+3 dB)</td>
<td>SDRs</td>
<td>4.0 (7.2)</td>
<td>4.2 (5.9)</td>
<td>3.2 (9.6)</td>
</tr>
<tr>
<td></td>
<td>SDRm</td>
<td>10.8 (4.5)</td>
<td>11.1 (3.5)</td>
<td>8.4 (5.7)</td>
</tr>
<tr>
<td>Male (-3 dB)</td>
<td>SDRs</td>
<td>0.1 (4.5)</td>
<td>1.5 (4.4)</td>
<td>-2.9 (4.4)</td>
</tr>
<tr>
<td></td>
<td>SDRm</td>
<td>13.1 (8.3)</td>
<td>14.9 (8.6)</td>
<td>8.5 (7.2)</td>
</tr>
<tr>
<td>Female (+3 dB)</td>
<td>SDRs</td>
<td>5.0 (8.1)</td>
<td>5.7 (7.3)</td>
<td>3.2 (8.0)</td>
</tr>
<tr>
<td></td>
<td>SDRm</td>
<td>9.6 (4.5)</td>
<td>10.7 (4.3)</td>
<td>7.3 (4.0)</td>
</tr>
<tr>
<td>Female (-3 dB)</td>
<td>SDRs</td>
<td>0.4 (4.6)</td>
<td>1.9 (5.0)</td>
<td>-2.0 (3.3)</td>
</tr>
<tr>
<td></td>
<td>SDRm</td>
<td>11.4 (7.9)</td>
<td>13.5 (8.8)</td>
<td>8.5 (6.1)</td>
</tr>
</tbody>
</table>

Speech / music source average SDR (dB) (SDRs / SDRm) computed on full-length sources and on segments of speech presence only (in braces).
Application to single channel speech / music separation

Audio examples

S-HMM (monophonic)

S-HMM (monophonic)

IS-NMF (polyphonic)

IS-NMF (polyphonic)

IS-NMF (polyphonic)

19 October 2009
Outline

- State-of-the-art models
- Motivation
- Factorial scaled hidden Markov model
- Inference algorithms
- Application to single-channel speech / music separation
- Conclusion
Conclusion

- Approach generalizing several existing models
- Modeling having the most credible physical motivation leads to the best separation results (SDR)

Further work

- Imagine other configurations of FS-HMM and apply it to other problems (e.g., music transcription)
- Speed up inference algorithms (e.g., via variational approximations)
- Extend to multichannel case (e.g., in line with [Ozerov & Févotte 2010]): quite straightforward
Thank you!


