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

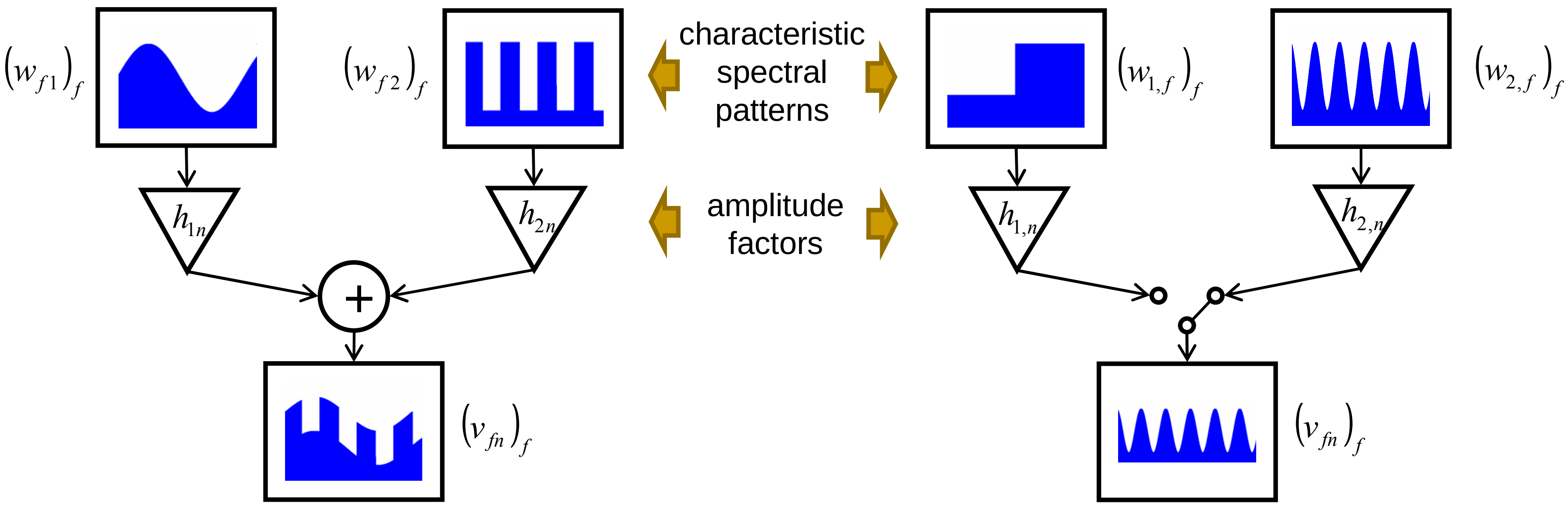
$$\boxed{X_{fn} \sim N_c\left(0, v_{fn}\right)} \qquad V = \left(v_{fn}\right)_{fn}$$

structured matrix

- 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]
- Gaussian scaled mixture model (GSMM) [Benaroya *et al* 2006]



- Suitable for polyphonic signals
- Suitable for monophonic signals

State-of-the-art models

- IS-NMF

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

- Summation of variances

- GSMM

$$p_{X_{fn}}(x) = \sum_{j=1}^{J} c_j N_c\left(x; 0, w_{j,f}h_{j,n}\right)$$

- 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])

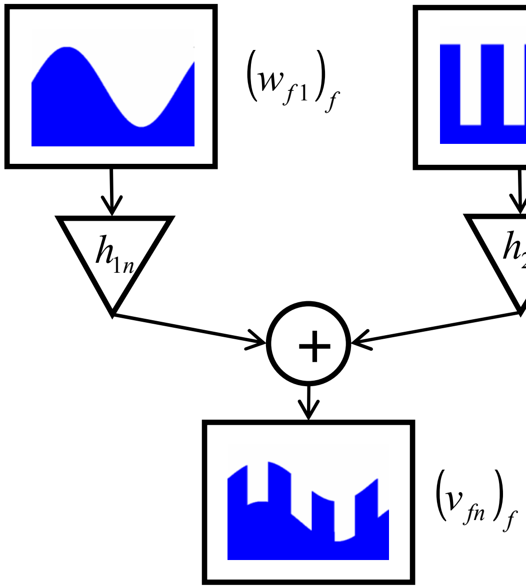
Outline

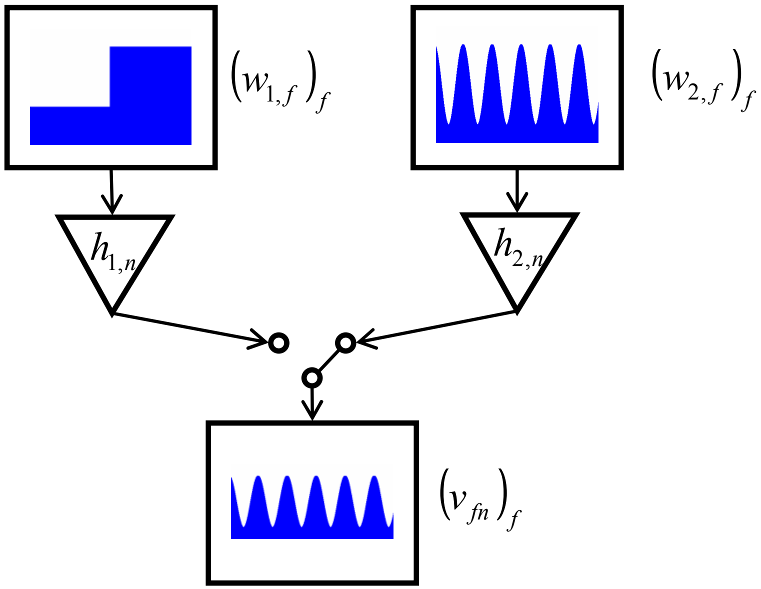
- State-of-the-art models
- Motivation
- Factorial scaled hidden Markov model
- Inference algorithms
- Application to single-channel speech / music separation
- Conclusion

Factorial scaled Hidden Markov model

- Mother (IS-NMF)

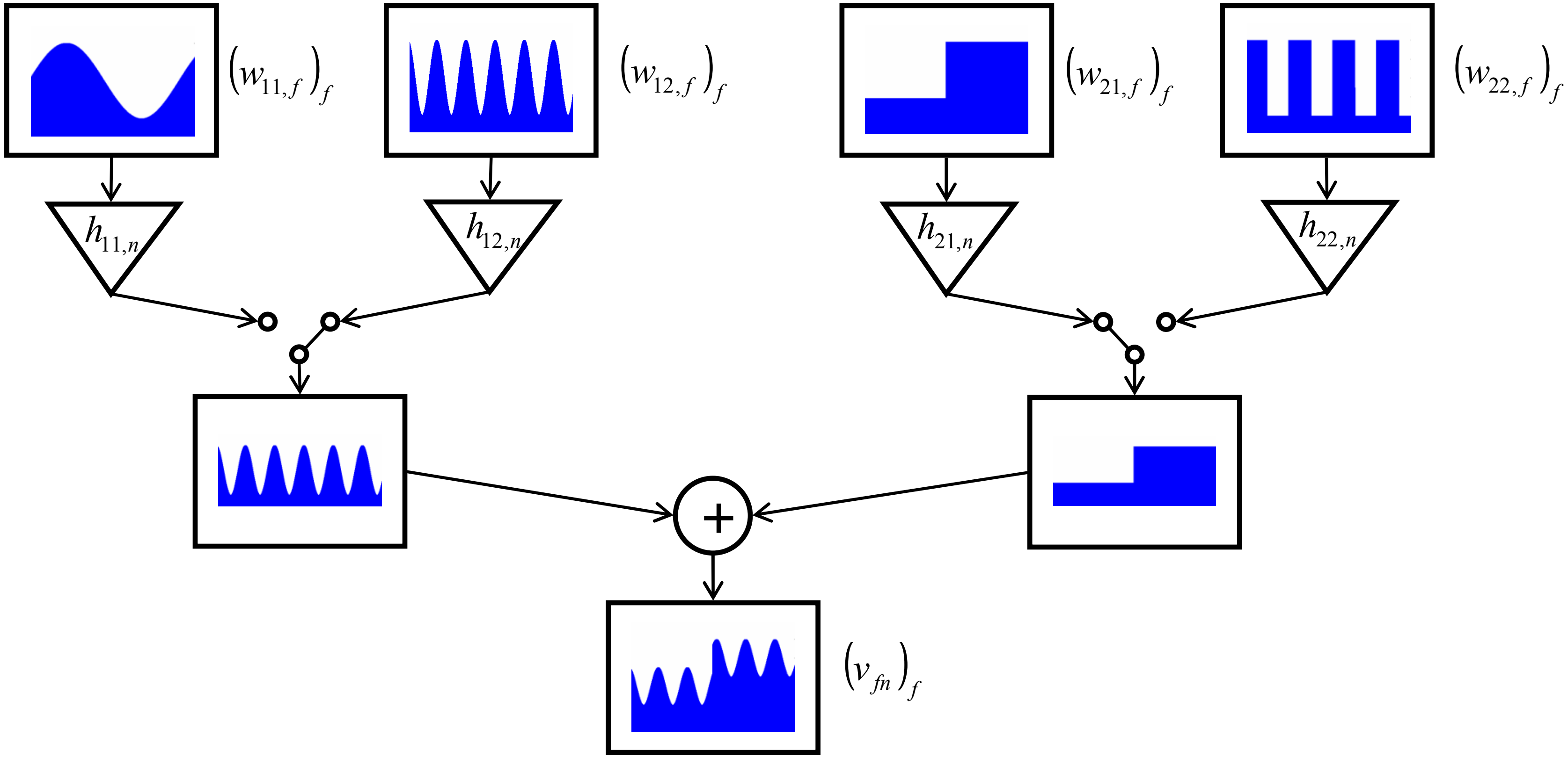
- Father (GSMM)





Factorial scaled Hidden Markov model

- Baby (Factorial scaled GMM or HMM)



Factorial scaled Hidden Markov model

Mother (IS-NMF)

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

Father (GSMM)

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

Baby 1 (FS-GMM)

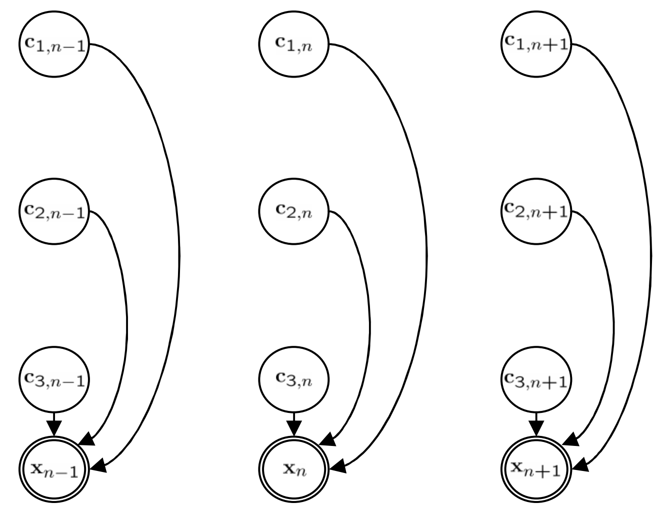
$$p_{X_{fn}}(x) = \left[\sum_{i_1=1}^{J_1} \alpha_i N_c\left(x; 0, w_{i_1,f1} h_{i_1,1n}\right)\right] \otimes \ldots \otimes \left[\sum_{i_K=1}^{J_K} \alpha_i N_c\left(x; 0, w_{i_K,fK} h_{i_K,Kn}\right)\right]$$

Baby 2 (FS-HMM)

We add temporal dependencies between states:
first order Markov chain

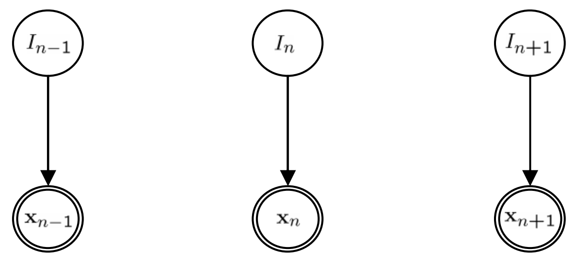
Factorial scaled Hidden Markov model

Mother (IS-NMF)



$$\mathbf{c}_{k,n} \sim \mathcal{N}_c(0, h_{kn}\mathrm{diag}(\mathbf{w}_k))$$

Father (GSMM)



$$\mathbf{x}_n \sim \mathcal{N}_c(0, h_{i,n}\mathrm{diag}(\mathbf{w}_i))$$

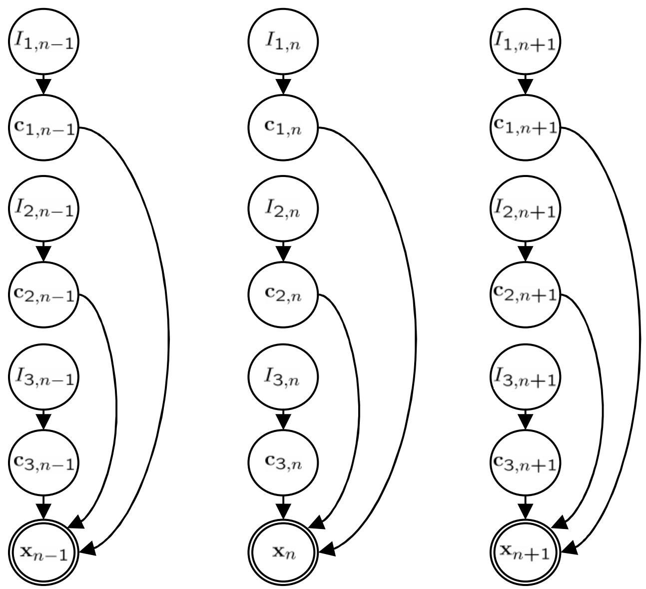
given $I_n = i$

$\mathbf{x}_n \in \mathbb{C}^F$

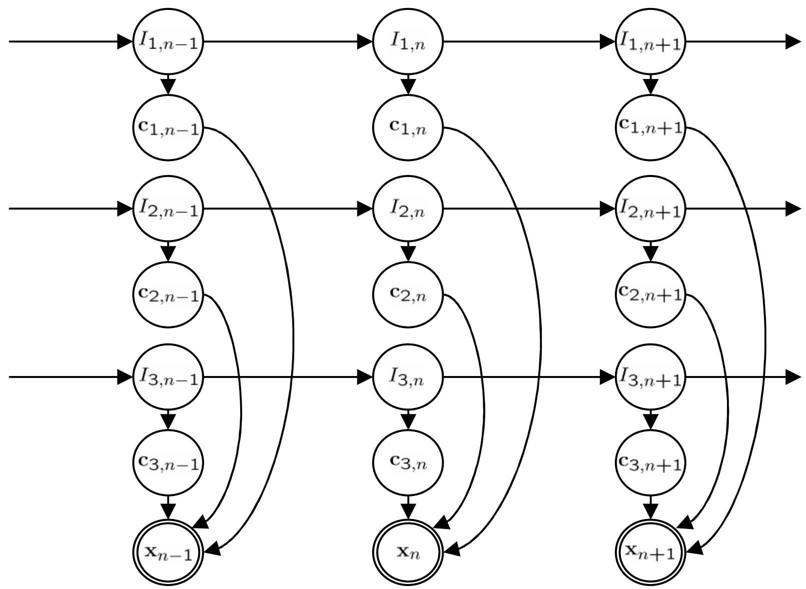
$\mathbf{c}_{k,n} \in \mathbb{C}^F$

Factorial scaled Hidden Markov model

Baby 1 (FS-GMM)



Baby 2 (FS-HMM)



$\mathbf{c}_{k,n} \sim \mathcal{N}_c(0, h_{i,kn}\mathrm{diag}(\mathbf{w}_{i,k}))$ given $I_{k,n} = i$

Outline

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

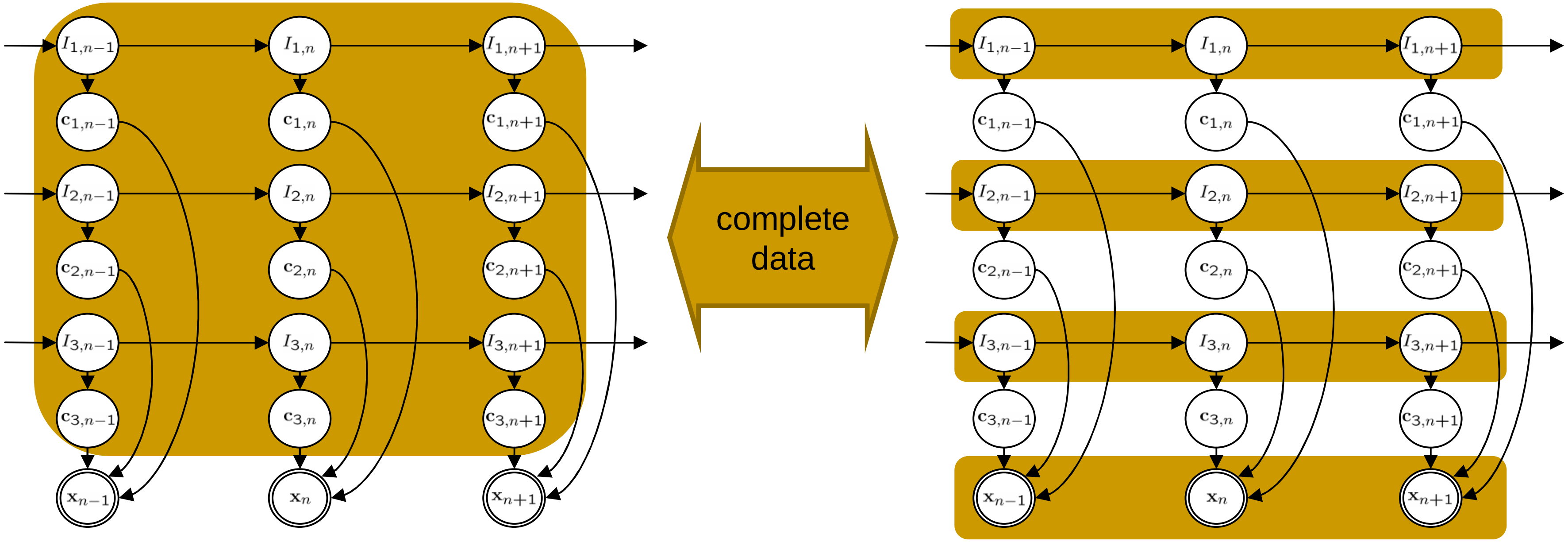
Inference algorithms

- Two Generalized EM (GEM) algorithms
- **EM algorithm**
  - Full complete data set:

$$\boxed{\mathcal{Y} = \{\mathbf{C}, \mathbf{I}\}}$$

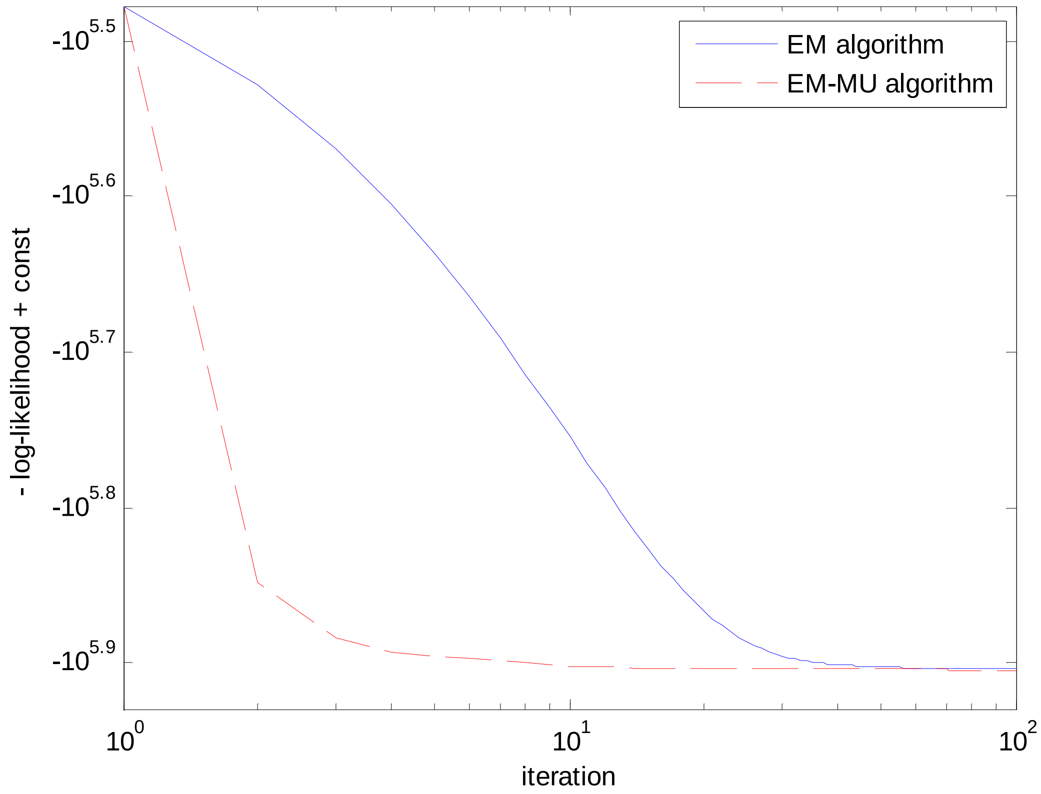
- **EM-MU algorithm**
  - Reduced complete data set:

$$\boxed{\mathcal{Z} = \{\mathbf{X}, \mathbf{I}\}}$$



Inference algorithms

- Convergence speeds



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

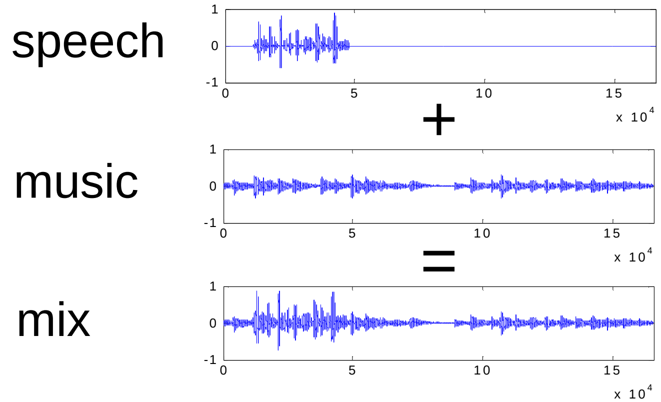
mix speech music

$$\mathbf{x}_n = \mathbf{s}_n + \mathbf{m}_n, \qquad \mathbf{s}_n = \sum_{k=1}^{K_s} \mathbf{c}_{k,n}, \qquad \mathbf{m}_n = \sum_{k=K_s+1}^{K_s+K_m} \mathbf{c}_{k,n},$$

- Tested model (FS-HMM) configurations :
  - **Mono. speech** / **Mono. music** (S-HMM / S-HMM):
    - Ks = Km = 1 (K = 2), J1 = 16, J2 = 8
  - **Mono. speech** / **Poly. music** (S-HMM / IS-NMF):
    - Ks = 1, Km = 8 (K = 9), J1 = 16, Jk = 1 (k > 1)
  - **Poly. speech** / **Poly. music** (IS-NMF / IS-NMF):
    - Ks = 16, Km = 8 (K = 24), Jk = 1 (k = 1, …, K)

Application to single channel speech / music separation

- Data



TIMIT (10 male speakers / 10 female speakers)

10 music 1 min. experts

- 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)

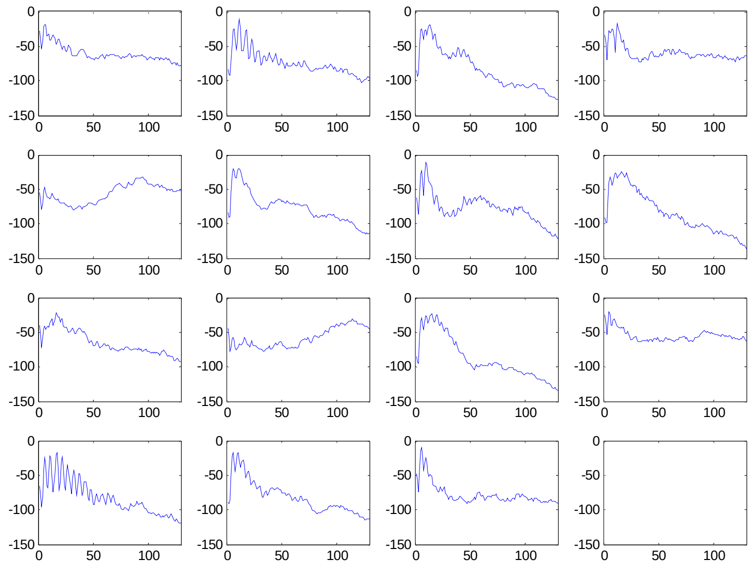
Application to single channel speech / music separation

- LEGO-like (e.g., modular framework)

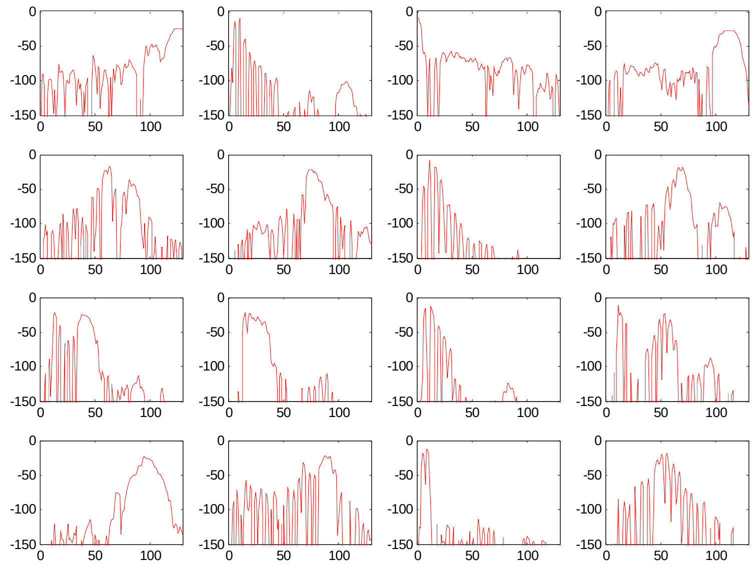
Application to single channel speech / music separation

- Female speech model spectral patterns

FS-HMM
(monophonic)



IS-NMF
(polyphonic)



Application to single channel speech / music separation

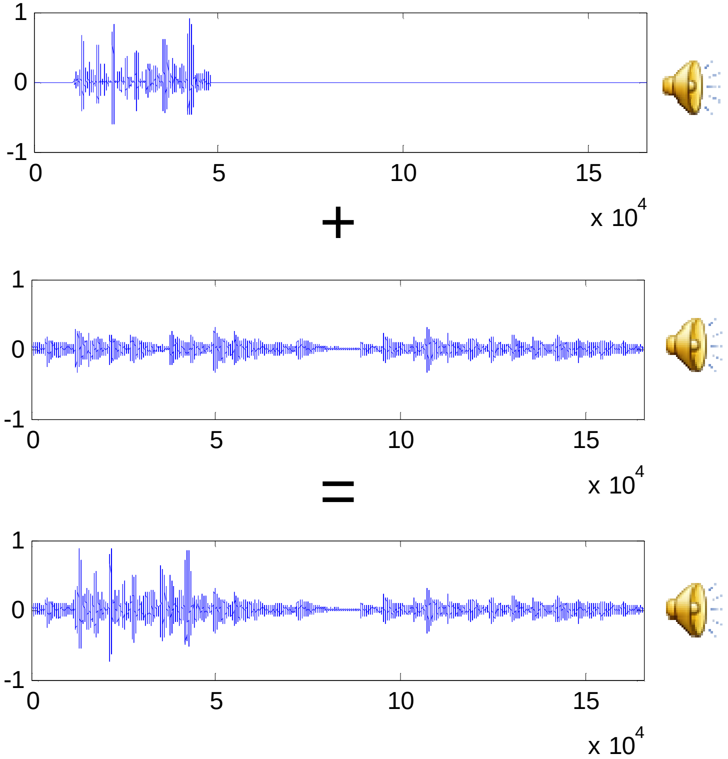
- Numerical results

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

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)

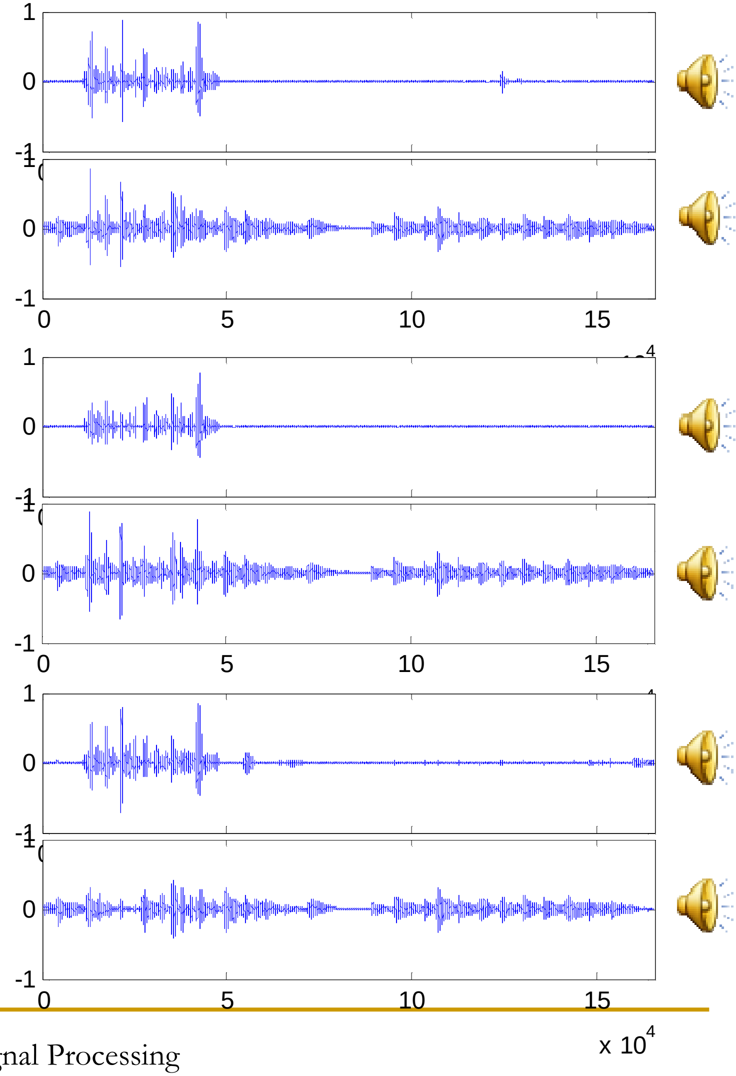
S-HMM (monophonic)

S-HMM (monophonic)

IS-NMF (polyphonic)

IS-NMF (polyphonic)

IS-NMF (polyphonic)



Outline

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

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!

References

- [Benaroya *et al* 2003] L. Benaroya, R. Gribonval, and F. Bimbot, “Non negative sparse representation for Wiener based source separation with a single sensor,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’03)*, Hong Kong, 2003, pp. 613–616.
- [Benaroya *et al* 2006] L. Benaroya, F. Bimbot, and R. Gribonval, “Audio source separation with a single sensor,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 191–199, 2006.
- [Bertin *et al* 2007] N. Bertin, R. Badeau, G. Richard, “Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark,” *International Conference on Audio, Speech and Signal Processing (ICASSP)*, Honolulu, 17-20 avril 2007.
- [Févotte *et al* 2009] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis,” *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.
- [Ozerov & Févotte 2010] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Audio, Speech and Lang. Proc.* special issue on Signal Models and Representations of Musical and Environmental Sounds , vol. 18, no. 1, Jan 2010 (*to appear*).