Factorial scaled hidden Markov model for polyphonic audio representation and source separation

Alexey Ozerov, Cédric Févotte, Maurice Charbit

FACTORIAL SCALED HIDDEN MARKOV MODEL FOR POLYPHONIC AUDIO REPRESENTATION AND SOURCE SEPARATION

Alexey Ozerov [1,], Cédric Févotte [2] and Maurice Charbit [1]*

[1]Institut Telecom, Telecom ParisTech, CNRS LTCI [2]CNRS LTCI, Telecom ParisTech
37-39, rue Dareau, 75014 Paris, France
{alexey.ozerov,cedric.fevotte,maurice.charbit}@telecom-paristech.fr

ABSTRACT

We present a new probabilistic model for polyphonic audio termed Factorial Scaled Hidden Markov Model (FS-HMM), which generalizes several existing models, notably the Gaussian scaled mixture model and the Itakura-Saito Nonnegative Matrix Factorization (NMF) model. We describe two expectation-maximization (EM) algorithms for maximum likelihood estimation, which differ by the choice of complete data set. The second EM algorithm, based on a reduced complete data set and multiplicative updates inspired from NMF methodology, exhibits much faster convergence. We consider the FS-HMM in different configurations for the difficult problem of speech / music separation from a single channel and report satisfying results.

Index Terms— Factorial hidden Markov model, Gaussian scaled mixture models, nonnegative matrix factorization, expectation-maximization algorithm, audio source separation.

1. INTRODUCTION

Many advanced probabilistic models for polyphonic audio have been proposed for source separation [1, 2, 3, 4] and semantic information retrieval (e.g., music transcription [5]). There is a close relation between source separation and semantic information retrieval in the sense that the outcome of one task can help the other. For example, singing voice separation is used to improve singer identification in [6] and, conversely, main melody line estimation is used to improve source separation in [7]. Therefore, it seems promising to perform these tasks jointly (e.g., models from [3] and [5] allow joint source separation and music transcription), thus exploiting the information from both acoustic and semantic levels in an optimal way. Such models should be generative so as to allow source reconstruction and include a semantic level, expressed for example in terms of some hidden states.

In this work we present a new generic model with the above-mentioned properties. We consider the family of models where the Short-Time Fourier Transform (STFT) of the audio signal is taken as realizations of zero-mean proper Gaussian multivariate random variables (possibly conditionally on some hidden state variables) [1, 2, 4, 8]. Such models have proven efficient for source separation; they allow in particular straightforward source estimation via (adaptive) Wiener filtering.

Benaroya *et al.* [2] model each source STFT by a Gaussian Mixture Model (GMM) modulated by a frame-dependent amplitude parameter accounting for nonstationarity, leading to the Gaussian Scaled Mixture Model (GSMM). Another model, better suited to polyphony, is proposed in [1] and was linked to Nonnegative Matrix Factorization (NMF) with the Itakura-Saito (IS) divergence by Févotte *et al.* [8]. This model takes each source STFT to be a sum of several elementary Gaussian components with fixed spectrum amplitude-modulated in frame. In short, the GSMM implicitly assumes the source to be monophonic with many possible states while the IS-NMF model assumes the source to be polyphonic, i.e., a sum of many elementary components with simple spectral signature.[1]

The aim of this paper is to bridge between the two models and introduce a general hybrid model that encompasses both the GSMM and NMF model. Moreover we incorporate time-persistence in the model through Markov modeling. In essence, our model assumes the observed composite signal to be a sum of independent components each modeled a Hidden Markov Model (HMM) with diagonal covariance matrices amplitude-modulated in frame. Our model, termed *Factorial Scaled Hidden Markov Model (FS-HMM)*, is described in details in Section 2. Then we describe in Section 3 two Expectation-Maximization (EM) algorithms for ML estimation. In Section 4 the convergence speed of the algorithms is compared, and the FS-HMM is applied in different configurations to the problem of speech / music separation from a single channel. Finally, conclusions are drawn in Section 5.

2. FACTORIAL SCALED HIDDEN MARKOV MODEL

Let $\mathbf{x}_n = [x_{fn}]_{f=1}^F$ be the complex-valued STFT of some audio data, termed *observation* ($f = 1, \dots, F$ is a frequency bin index, $n = 1, \dots, N$ is a time frame index). The following composite model is considered:

$$\mathbf{x}_n = \sum\nolimits_{k=1}^{K} \mathbf{c}_{k,n}, \tag{1}$$

where the complex-valued vectors $\mathbf{c}_{k,n} = [c_{k,fn}]_{f=1}^F$ are referred to as *components* and assumed mutually independent (i.e., over k). Let $I_{k,n}$ be J_k-states discrete random variables (r.v.) independent over k. We assume for each component $\mathbf{c}_{k,n}$ the following mixture model

$$\mathbf{c}_{k,n} = \sum\nolimits_{i=1}^{J_k} \mathbf{u}_{ki,n} \mathbb{1}(I_{k,n} = i), \tag{2}$$

where the *sub-components* $\mathbf{u}_{ki,n} = [u_{ki,fn}]_{f=1}^F$ are assumed independent (over component k, state i and time n) proper complex

[*] A. Ozerov is now with IRISA, Metiss Group, Rennes, France.

This work was supported in part by the French ANR project SARAH.

[1] In other words, with GSMM a summation takes place in the probability density function (pdf) of the source STFT while with IS-NMF the summation takes place in the STFT domain.

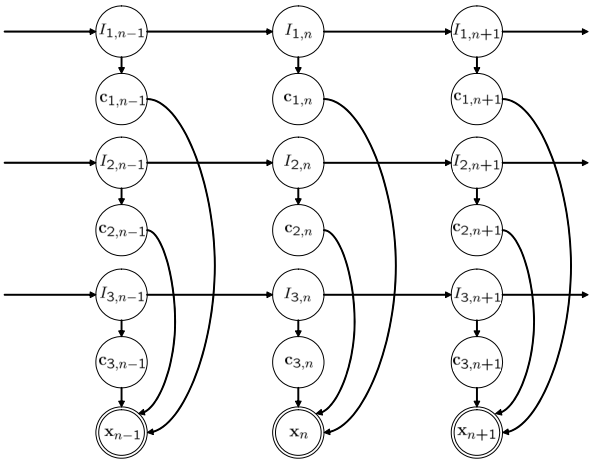

Figure 1: Bayesian network representing an FS-HMM with $K = 3$ components ($\mathbf{c}_{k,n} \sim \mathcal{N}_c(0, h_{ki,n}\text{diag}\,(\mathbf{w}_{ki}))$, given $I_{k,n} = i$).

zero-mean Gaussian vectors with structured diagonal covariance matrices

$$\mathbf{\Sigma}_{ki,n} = h_{ki,n}\text{diag}\,(\mathbf{w}_{ki})\,, \tag{3}$$

where the vector $\mathbf{w}_{ki} = [w_{ki,f}]_{f=1}^{F}$ ($w_{ki,f} > 0$) and the scalar $h_{ki,n} > 0$ represent respectively the characteristic spectral pattern and amplitude factor of $\mathbf{c}_{k,n}$ as in state i. Assuming discrete r.v. $I_{k,n}$ independent over time n, the proposed model reduces to the GSMM [2] [2] for $K = 1$ and to the IS-NMF model [8] for $J_k = 1$ ($k = 1, \ldots, K$), thus generalizing both models. Moreover, we introduce time dependencies in the underlying discrete process, by assuming $I_{k,n}$ to be a first-order Markov chain specified by transition probabilities:

$$a_{kij} = \mathbb{P}(I_{k,n+1} = j | I_{k,n} = i). \tag{4}$$

Altogether the model parameters can be written as:

$$\begin{aligned}\mathbf{\Theta} &= \{\mathcal{A}_k, \mathcal{W}_k, \mathcal{H}_k\}_{k=1}^{K} \\ &\stackrel{\text{def}}{=} \left\{\{a_{kij}\}_{i,j=1}^{J_k,J_k}, \{w_{ki,f}\}_{i,f=1}^{J_k,F}, \{h_{ki,n}\}_{i,n=1}^{J_k,N}\right\}_{k=1}^{K}\end{aligned} \tag{5}$$

The term "factorial" in FS-HMM refers to the fact that the observation $\mathbf{x}_n$ is Gaussian conditionally on so-called *factorial state r.v.* $\mathcal{I}_n = (I_{1,n}, I_{2,n}, \ldots, I_{K,n})$ taking values in so-called *factorial state space* $\mathcal{Q} = \{\mathbb{I} = (i_1, i_2, \ldots, i_K)\}_{i_1,i_2,\ldots,i_K=1}^{J_1,J_2,\ldots,J_K}$ that can be "factorized" into individual state spaces $\{i_k\}_{i_k=1}^{J_k}$ [9]. Figure 1 represents the Bayesian network corresponding to the FS-HMM.

3. ML PARAMETERS ESTIMATION

Given the observation $\mathbf{X} = [\mathbf{x}_{fn}]_{f,n}$, the ML estimate of $\mathbf{\Theta}$ can be computed using an EM algorithm. We here consider two EM algorithms. [3] The first algorithm, referred to as *EM algorithm*, is based on complete data $\mathcal{Y} = \{\mathbf{C}, \mathbf{I}\}$, where $\mathbf{C} = [c_{k,fn}]_{k,f,n}$ is the $K \times F \times N$ complex-valued array of components and $\mathbf{I} = [I_{k,n}]_{k,n}$ is the $K \times N$ array of state variables. The second algorithm, referred to as *EM-MU algorithm*, is based on the reduced complete data set $\mathcal{Z} = \{\mathbf{X}, \mathbf{I}\}$, and its M step is optimized using multiplicative updates (MU). In the following, for sake of brevity, we only sketch the algorithms and give the final update rules, which may be derived with no specific complication.

3.1. EM algorithm

It can be shown that the pdf $p(\mathbf{C}, \mathbf{I}|\mathbf{\Theta})$ of the complete data $\mathcal{Y}$ belongs to an *exponential family* [10] with the *natural sufficient statistics* $\mathbf{T}_1 = \left\{\{\mathbf{t}^0_{ki,n}\}_{k,i,n=1}^{K,J_k,N}, \{\mathbf{t}^0_{kij,n}\}_{k,i,j,n=1}^{K,J_k,J_k,N}, \{\mathbf{t}^2_{ki,fn}\}_{k,i,n,f=1}^{K,J_k,N,F}\right\}$, where:

$$\mathbf{t}^0_{ki,n} = \mathbb{1}(I_{k,n} = i), \tag{6}$$

$$\mathbf{t}^0_{kij,n} = \mathbb{1}(I_{k,n} = i, I_{k,n+1} = j), \tag{7}$$

$$\mathbf{t}^2_{ki,fn} = |c_{k,fn}|^2 \mathbb{1}(I_{k,n} = i). \tag{8}$$

The corresponding EM algorithm is summarized below, where (l) denotes parameter estimates from l-th iteration.

E step - Compute the conditional expectation $\widehat{\mathbf{T}}_1 = \mathbb{E}\left[\mathbf{T}_1|\mathbf{X}; \mathbf{\Theta}^{(l)}\right]$ using a forward-backward procedure [11] in the factorial state space $\mathcal{Q}$.

M step - Update model parameters:

$$a^{(l+1)}_{kij} = \sum\nolimits_{n=1}^{N-1} \hat{\mathbf{t}}^0_{kij,n} \Big/ \sum\nolimits_{n=1}^{N-1} \hat{\mathbf{t}}^0_{k,i,n}\,, \tag{9}$$

$$w^{(l+1)}_{ki,f} = \sum\nolimits_{n} \left(\hat{\mathbf{t}}^2_{ki,fn} \Big/ h^{(l)}_{ki,n}\right) \Big/ \sum\nolimits_{n} \hat{\mathbf{t}}^0_{ki,n}\,, \tag{10}$$

$$h^{(l+1)}_{ki,n} = \sum\nolimits_{f} \left(\hat{\mathbf{t}}^2_{ki,fn} \Big/ w^{(l+1)}_{ki,f}\right) / (F \cdot \hat{\mathbf{t}}^0_{ki,n}). \tag{11}$$

3.2. EM-MU algorithm

We here take the complete data as $\mathcal{Z} = \{\mathbf{X}, \mathbf{I}\}$. In this case the complete data pdf $p(\mathbf{X}, \mathbf{I}|\mathbf{\Theta})$ also belongs to an exponential family with the natural sufficient statistics $\mathbf{T}_2 = \left\{\{\mathbf{t}^0_{kij,n}\}_{k,i,j,n=1}^{K,J_k,J_k,N}, \{\{\mathbf{t}^0_{\mathbb{I},n}\}_{\mathbb{I}\in\mathcal{Q}}\}_{n=1}^{N}, \{\{\mathbf{t}^2_{\mathbb{I},fn}\}_{\mathbb{I}\in\mathcal{Q}}\}_{f,n=1}^{F,N}\right\}$, where $\mathbf{t}^0_{kij,n}$ is defined by (7) and

$$\mathbf{t}^0_{\mathbb{I},n} = \mathbb{1}(\mathcal{I}_n = \mathbb{I}), \tag{12}$$

$$\mathbf{t}^2_{\mathbb{I},fn} = |x_{fn}|^2 \mathbb{1}(\mathcal{I}_n = \mathbb{I}). \tag{13}$$

Due to the reduced, less informative, complete data set the M step is more difficult to solve. However, the complete data log-likelihood can be brought down to a somehow more standard nonnegative decomposition problem with the IS divergence $d_{IS}(x|y) = x/y - \log(x/y) - 1$:

$$\begin{aligned}&\log p(\mathbf{X}, \mathbf{I}|\mathbf{\Theta}) = \\ &-\sum\nolimits_{\mathbb{I}\in\mathcal{Q}} \sum\nolimits_{f,n} d_{IS}\left(|x_{fn}|^2 \Big| \sum\nolimits_{k} w_{ki_k,f} h_{ki_k,n}\right) \mathbf{t}^0_{\mathbb{I},n} + c,\end{aligned}$$

where c is a constant term independent on $w_{ki,f}$ and $h_{ki,n}$. Thus, we can apply MU rules from [8], leading to the EM-MU algorithm summarized below. It was always observed in practice that with these update rules the criterion is monotonically non-decreasing. However, in our best knowledge, it was not yet proven.

[2] In [2] a partial case of factorial GSMM with $K = 2$ is also addressed, and a method for inference of amplitude factors is given. Our framework is more general since we consider any K and we address the inference of both amplitude factors $h_{ki,n}$ and spectral patterns $\mathbf{w}_{ki}$ from the observation.

[3] Both EM algorithms considered here are strictly speaking only generalized EM (GEM) algorithms because their M step does not ensure the auxiliary function to be maximized, but only to be non-decreasing.

E step: Compute the conditional expectation $\widehat{\mathbf{T}}_2 = \mathbb{E}\left[\mathbf{T}_2|\mathbf{X};\boldsymbol{\Theta}^{(l)}\right]$ using a forward-backward procedure [11].

M step: Update a_{kij} using Eq. (9) and update $w_{ki,f}$ and $h_{ki,n}$ as

$$\mathbf{w}_{ki}^{(l+1)} = \mathbf{w}_{ki}^{(l)} \frac{\mathbf{V}.\left\langle \left(\mathbf{W}^{(l)}\mathbf{H}^{(l)}\right)^{.-2}\right\rangle_{ki}\left\{\mathbf{h}_{ki}^{(l)}\right\}^T}{\left\langle \left(\mathbf{W}^{(l)}\mathbf{H}^{(l)}\right)^{.-1}\right\rangle_{ki}\left\{\mathbf{h}_{ki}^{(l)}\right\}^T},$$

$$\mathbf{h}_{ki}^{(l+1)} = \mathbf{h}_{ki}^{(l)} \frac{\left\{\mathbf{w}_{ki}^{(l+1)}\right\}^T \mathbf{V}.\left\langle \left(\mathbf{W}^{(l+1)}\mathbf{H}^{(l)}\right)^{.-2}\right\rangle_{ki}}{\left\{\mathbf{w}_{ki}^{(l+1)}\right\}^T \left\langle \left(\mathbf{W}^{(l+1)}\mathbf{H}^{(l)}\right)^{.-1}\right\rangle_{ki}},$$

where $\mathbf{V}$ is the $F \times N$ matrix with elements $v_{fn} = |x_{fn}|^2$, $\mathbf{w}_{ki}$ is the $F \times 1$ column vector with elements $w_{ki,f}$, $\mathbf{h}_{ki}$ is the $1 \times N$ row vector with elements $h_{ki,n}$, "." indicates element-wise matrix operations, and (for $p = 1, 2$)

$$\left\langle (\mathbf{WH})^{.-p}\right\rangle_{ki} = \sum\nolimits_{\mathbb{I}\in\mathcal{Q}\cap\{i_k=i\}} (\mathbf{W}_{\mathbb{I}}\mathbf{H}_{\mathbb{I}})^{.-p}.(\mathbf{1}_{F\times 1}\mathcal{G}_{\mathbb{I}}),$$

where $\mathbf{W}_{\mathbb{I}}$ (resp. $\mathbf{H}_{\mathbb{I}}$) is the $F \times K$ (resp. $K \times N$) matrix with columns $\mathbf{w}_{ki_k}$ (resp. rows $\mathbf{h}_{ki_k}$), $\mathbf{1}_{F\times 1}$ is a $F \times 1$ column vector of ones, and $\mathcal{G}_{\mathbb{I}}$ is the $1 \times N$ vector with elements $\hat{\mathbf{t}}^0_{\mathbb{I},n}$.

4. RESULTS

All the audio signals used in this experimental part are mono and sampled at 11025 Hz. The STFT is computed using a half-overlapping 23 ms (256 samples) length Hann window.

4.1. Convergence speed

We used the STFT of an arbitrary 8 second length music signal and ran 100 iterations of both EM and EM-MU algorithms with $K = 3$ components, $J_k = 3$ states per component, and using the same random parameters initialization. Negative log-likelihoods of both algorithms are plotted on Figure 2 as functions of the iteration number. We see that the EM-MU algorithm converges much faster than the EM algorithm, a consequence of the less informative complete data set used in the former. As such, we only consider the EM-MU algorithm in the following experiments.

4.2. Application to single-channel speech / music separation

We now apply FS-HMM to the challenging problem of separating a speech signal from a background music [12]. We assume that the set $\mathcal{K} = \{1, 2, \ldots, K\}$ of FS-HMM component indices is split onto two disjoint subsets $\mathcal{K}_s = \{1, \ldots, K_s\}$ and $\mathcal{K}_m = \{K_s + 1, \ldots, K_s + K_m\}$, corresponding to indices of speech and music source components. In other words, Eq. (1) can be rewritten as:

$$\mathbf{x}_n = \mathbf{s}_n + \mathbf{m}_n, \quad \mathbf{s}_n = \sum\nolimits_{k\in\mathcal{K}_s} \mathbf{c}_{k,n}, \quad \mathbf{m}_n = \sum\nolimits_{k\in\mathcal{K}_m} \mathbf{c}_{k,n},$$

where $\mathbf{s}_n$ and $\mathbf{m}_n$ are the STFT frames of speech and music sources, respectively.

Single-channel source separation is a difficult task, and fully blind separation approaches usually do not work, due to the lack of *a priori* knowledge about the sources. We here consider a supervised approach, where some FS-HMM parameters corresponding to speech source are pre-trained from training data. We more precisely consider the following setting:

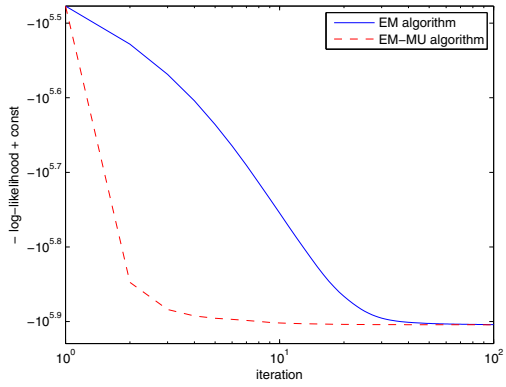

Figure 2: Convergence of EM algorithm vs. EM-MU algorithm. In our MATLAB implementation, 100 iterations of EM (resp. EM-MU) algorithm take about 170 sec (resp. 280 sec) for this example.

- **Training:** Estimate speech source FS-HMM $\boldsymbol{\Theta}_s = \{\mathcal{A}_k, \mathcal{W}_k, \mathcal{H}_k\}_{k=1}^{K_s}$ (see Eq. (5)) from speech training data using 100 iterations of EM-MU algorithm and random parameters initialization.
- **Test:** Given the STFT of the mixture signal $\mathbf{x}_n = \mathbf{s}_n + \mathbf{m}_n$ $(n = 1, \ldots, N)$, perform the following steps:
  - *Initialization:* Initialize transition probabilities $\mathcal{A}_k$ and spectral patterns $\mathcal{W}_k$ of speech source (i.e., for $k \in \mathcal{K}_s$) with pre-trained values, and initialize all other parameters of $\boldsymbol{\Theta} = \{\mathcal{A}_k, \mathcal{W}_k, \mathcal{H}_k\}_{k=1}^{K}$ randomly.
  - *Inference:* Run 100 iterations of the EM-MU algorithm, while keeping pre-trained parameters $\{\mathcal{A}_k, \mathcal{W}_k\}_{k=1}^{K_s}$ fixed.
  - *Separation:* Compute Minimum Mean Square Error (MMSE) speech STFT estimate $\hat{s}_{fn} = \mathbb{E}\left[s_{fn}|\mathbf{X};\boldsymbol{\Theta}\right]$ (and $\hat{m}_{fn}$ similarly). Finally, reconstruct time-domain source estimates with inverse-STFT. This reconstruction is conservative (i.e., $x_{fn} = \hat{s}_{fn} + \hat{m}_{fn}$) in both STFT and time domains.

Speaker's gender is assumed to be known, and gender-specific pre-trained speech parameters are used. To train male (resp. female) speech parameters we used 10 sentences from 10 different male (resp. female) speakers randomly selected from TIMIT database training part. As for music sources, we used 10 (15 s length) music samples randomly selected from 10 music pieces. As for speech sources, we used 20 sentences from 20 different speakers (10 male and 10 female speakers) randomly selected from the TIMIT database evaluation part. These (1 to 6 s length) sentences were randomly placed into 15 s intervals, and padded with zeros, thus creating 15 s length speech sources. Finally, 10 male and 10 female speech sources were mixed with corresponding 10 music sources using two levels of Speech to Music Ratio (SMR), namely +3 and -3 dB. We have tested the following different configurations of FS-HMM parameters, corresponding to modeling speech and music sources by either a Scaled HMM (S-HMM) or an NMF model:

- **S-HMM / S-HMM**: $K_s = K_m = 1$ $(K = 2)$, $J_1 = 16$, $J_2 = 8$. Speech and music sources are modeled by S-HMMs

with 16 and 8 states, respectively. This configuration shares some common points with [4], except that S-HMMs are used instead of GMMs, and there is no need for (speech + music) / music segmentation.

- **S-HMM / NMF**: $K_s = 1$, $K_m = 8$ ($K = 9$), $J_1 = 16$, $J_k = 1 (k > 1)$. The speech source is modeled by an S-HMM with 16 states, and the music source is modeled by an NMF model with 8 components.
- **NMF / NMF**: $K_s = 16$, $K_m = 8$ ($K = 24$), $J_k = 1 (k = 1, \ldots, K)$. Speech and music sources are modeled by NMF models with 16 and 8 components, respectively. This configuration shares some common points with [7], except that the speech model is here pre-trained.

Table 1 summarizes the separation results in terms of average Source to Distortion Ratio (SDR) (see e.g., [4]) computed on full-length sources and on segments of speech presence only (in braces). We see that the S-HMM / NMF hybrid modeling leads consistently to the best average SDR computed on full-length sources. As a matter of fact, the S-HMM / NMF modeling is the best motivated by the physical nature of the sources. Indeed, a monophonic speech spectrum is better representable by a single scaled spectral pattern (S-HMM), while a polyphonic music spectrum is better represented by a sum of spectral patterns (NMF model). This result is also consistent with conclusions drawn by Blouet *et al.* [12]. From informal listening of the separated sources [4] we have noticed that the NMF / NMF modeling leads to the best preserved speech signal, but also leaves significant music interferences in the speech estimate. Though the S-HMM / NMF produced a rather corrupted speech estimate, the music is better suppressed everywhere in the speech estimate, and particularly in the parts where there is no speech.

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

Table 1: Speech / music source average SDR (dB) (SDRs / SDRm) computed on full-length sources and on segments of speech presence only (in braces).

5. CONCLUSION

We have introduced a novel model for polyphonic audio. Our model, FS-HMM, is a generalization of existing models built on Gaussian assumptions. We have designed two EM algorithms for ML estimation of the parameters, and confirmed experimentally that one algorithm is much faster than the other, as explained from the different level of "completeness" of the chosen missing data in each case. The generality of the FS-HMM has allowed us to test several modeling strategies for single channel speech / music separation in a realistic setting. Experimental evaluation showed that the modeling having the most credible physical motivation leads indeed to the best separation results. Further research directions include applying FS-HMM to other source separation tasks and also to semantic information retrieval. Playing with model orders (i.e., K and J_k) and fixing *a priori* different parameters can lead to various blind and (semi-)supervised decomposition approaches. As the computational complexity of the proposed EM algorithms is of the order of $(J_1 \times \ldots \times J_K)^2$ it can become quickly computationally intractable when the number of states allotted to each component grows. As such another research direction is to consider variational approximations in the line of [9].

[4] Some separation examples are available at `http://perso.telecom-paristech.fr/~ozerov/demos.html#waspaa09`.

6. ACKNOWLEDGMENT

The authors would like to thank J.-L. Durrieu for useful discussions about EM-MU algorithm.

7. REFERENCES

[1] 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.

[2] 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.

[3] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 91–98, 2006.

[4] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 15, no. 5, pp. 1564–1578, July 2007.

[5] A. T. Cemgil, H. J. Kappen, and D. Barber, "A Generative Model for Music Transcription," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 679–694, March 2006.

[6] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in *International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[7] J.-L. Durrieu, G. Richard, and B. David, "An iterative approach to monaural musical mixture de-soloing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'09)*, 2009.

[8] 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.

[9] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, pp. 245–273, 1997.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[11] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[12] R. Blouet, G. Rapaport, I. Cohen, and C. Févotte, "Evaluation of several strategies for single sensor speech/music separation," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, Las Vegas, USA, Apr. 2008.