Generative Spectrogram Factorization Models for Polyphonic Piano Transcription

Paul H. Peeling, A. Taylan Cemgil, *Member, IEEE*, and Simon J. Godsill, *Member, IEEE*

***Abstract*—We introduce a framework for probabilistic generative models of time–frequency coefficients of audio signals, using a matrix factorization parametrization to jointly model spectral characteristics such as harmonicity and temporal activations and excitations. The models represent the observed data as the superposition of statistically independent sources, and we consider variance-based models used in source separation and intensity-based models for non-negative matrix factorization. We derive a generalized expectation-maximization algorithm for inferring the parameters of the model and then adapt this algorithm for the task of polyphonic transcription of music using labeled training data. The performance of the system is compared to that of existing discriminative and model-based approaches on a dataset of solo piano music.**

***Index Terms*—Frequency estimation, matrix decomposition, music information retrieval (MIR), spectral analysis, time–frequency analysis.**

I. Introduction

NUMEROUS authors have focused on the problem of the transcription of solo recordings of polyphonic piano music, using a wide variety of techniques and approaches. There is some growing consensus on suitable evaluation criteria to assess the performance of these systems, which is forming within the MIREX community,[1] particularly the "Multiple Fundamental Frequency Estimation and Tracking task." However, as a subset of these approaches, there also exist systems which are capable of performing multiple-pitch classification on individual time-localized frames of audio data, a task known as frame-level transcription. In a data-driven approach, frame-level transcription can be viewed as a preprocessing step, whereas in a Bayesian approach, the frame-level transcription is due to the signal source model, over which priors for the transitions of note pitches between frames can be introduced. Frame-level transcription can therefore be used to assess the performance in isolation of the source model in a music transcription system.

Manuscript received December 29, 2008; revised June 30, 2009. Current version published February 10, 2010. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/D03261X/1 entitled "Probabilistic Modeling of Musical Audio for Machine Listening." The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Paris Smaragdis.

P. H. Peeling and S. J. Godsill are with the Signal Processing and Communications Laboratory, Engineering Department, Cambridge University, Cambridge CB2 1PZ, U.K. (e-mail: php23@cam.ac.uk; sjg@eng.cam.ac.uk).

A. T. Cemgil is with the Department of Computer Engineering, Boğaziçi University, 34342 Bebek, Istanbul, Turkey (e-mail: taylan.cemgil@boun.edu.tr).

Digital Object Identifier 10.1109/TASL.2009.2029769

A useful comparative study of three varied approaches has been carried out by Poliner and Ellis in [1]. A dataset with ground-truth of polyphonic piano music has been provided to assess the performance of a support vector machine (SVM) classifier, [1], further improved with regards to generalization in [2], which is provided as an example of a discriminative based approach, having favorable performance in classification accuracy; a neural-network classifier [3], known as SONIC[2]; and an auditory-model based approach [4].

Generative models, which rely on a prior model of musical notes, which for example include [5]–[8], have not been comprehensively evaluated in such a framework, as Poliner and Ellis pursue the insight that a prior model based on harmonics was unnecessary for transcription.

Another class of techniques that have recently become popular for transcription are based on non-negative matrix factorization, factorizing a matrix of time–frequency coefficients into a codebook of spectral templates and an activation matrix from which the transcription can be inferred. These approaches do not typically have a prior model of musical notes, but this is readily learned by supplying training data. Bayesian approaches allow the inclusion of priors and more powerful inference techniques, and these have been applied in the field of polyphonic music transcription in [9]–[12].

The contribution of this paper is to extend the comparative study in [1] to include non-negative matrix factorization approaches. The difficulty in applying these approaches to classification is the joint problem of choosing the number of single rank matrices (sources) to perform the approximation, and labeling the activation matrix in terms of the active notes. However, by adopting a prior structure conditioned on the pitch and velocity of notes, and by adopting the generative interpretation of matrix factorization as the superposition of independent sources, we are able to address this in our inference scheme. We will show that transcription is a result, or by-product, of inferring the model parameters. Our emphasis will therefore be in designing suitable models for the spectrogram coefficients in polyphonic piano music and using transcription to assess the suitability of such models, rather than selecting the optimum hyperparameters in the prior for transcription performance.

The overview of the paper is as follows. In Section II, we describe the non-negative matrix factorization (NMF) model as applied to matrices of time–frequency coefficients (spectrograms). This section includes a general formulation of the model, and then two specific choices of signal model: first, the commonly used NMF divergence measure, which can be interpreted as a Poisson distribution with parametrized

[1]http://www.music-ir.org/mirex

[2]http://lgm.fri.uni-lj.si/SONIC

intensity; second, a source separation model using the normal distribution with zero mean and parametrized variance as the source model, and finally derives the expectation-maximization (EM) algorithm for finding the maximum *a posteriori* (MAP) estimate of the parameters. In Section III, we describe how the EM algorithm can be adapted to infer polyphonic frame-level transcription, and describe a particular prior structure that can be placed over the activation matrix. In Section IV, we compare the performance of the matrix factorization models to previously evaluated approaches, and in Section V we comment on the implications of the comparison and how the inference and prior structure can be improved further in the frame-level transcription setting.

II. Spectrogram Factorization Models

A. General Formulation

We construct a matrix $\mathbf{X} \in \mathbb{C}^{F \times K}$ of time–frequency coefficients, which is drawn from a probability distribution $p(\mathbf{X}|\mathbf{TV})$ parametrized by the product of a matrix of spectral templates $\mathbf{T} \in \mathbb{R}_+^{F \times N}$ and an excitation or activation matrix $\mathbf{V} \in \mathbb{R}_+^{N \times K}$. The matrix product can be viewed as separating the observation into N conditionally independent sources $\mathbf{S}_n, n = 1, \ldots, N$ where each source matrix of time–frequency coefficients is parametrized by the rank-one product $\mathbf{t}_n \mathbf{v}_n^\top$ of the nth column vector $\mathbf{t}_n$ of $\mathbf{T}$ and the nth row vector $\mathbf{v}_n$ of $\mathbf{V}$. Each individual source has the same probability distribution

$$\mathbf{S}_n \sim p\left(\mathbf{S}_n|\mathbf{t}_n\mathbf{v}_n^\top\right)$$

and so the observed matrix $\mathbf{X}$ is the superposition of the sources

$$\mathbf{X} = \sum_{n=1}^{N} \mathbf{S}_n \sim p\left(\mathbf{X}|\sum_{n=1}^{N} \mathbf{t}_n\mathbf{v}_n^\top\right) \sim p(\mathbf{X}|\mathbf{TV}). \quad (1)$$

The joint probability distribution of the observed matrix $\mathbf{X}$ and the sources $\mathbf{S}_n$ model can be equivalently expressed as

$$p(\mathbf{X}, \mathbf{S}_1, \ldots, \mathbf{S}_N|\mathbf{TV}) = \delta\left(\mathbf{X} - \sum_{n=1}^{N} \mathbf{S}_n\right) \prod_{n=1}^{N} p\left(\mathbf{S}_n|\mathbf{t}_n\mathbf{v}_n^\top\right)$$

which we will express more succinctly, by grouping the sources together $\mathbf{S} = \{\mathbf{S}_1, \ldots, \mathbf{S}_N\}$, as

$$p(\mathbf{X}, \mathbf{S}|\mathbf{TV}) = p(\mathbf{X}|\mathbf{S})\, p(\mathbf{S}|\mathbf{TV}).$$

The joint probability of the spectrogram factorization model is thus

$$p(\mathbf{X}, \mathbf{S}, \mathbf{T}, \mathbf{V}) = p(\mathbf{X}|\mathbf{S})\, p(\mathbf{S}|\mathbf{TV})\, p(\mathbf{T}, \mathbf{V}). \quad (2)$$

B. Expectation-Maximization Algorithm

For appropriate choices of probability distributions and conjugate priors, we can find a local maximum of the log likelihood of the generative spectrogram factorization model efficiently by the EM algorithm [13]. The log likelihood is approximated with an instrumental distribution $q(\mathbf{S})$ as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{T}, \mathbf{V}) &\equiv \log \sum_{\mathbf{S}} p(\mathbf{X}, \mathbf{S}, \mathbf{T}, \mathbf{V}) \\ &\geq \sum_{\mathbf{S}} q(\mathbf{S}) \log \frac{p(\mathbf{X}, \mathbf{S}, \mathbf{T}, \mathbf{V})}{q(\mathbf{S})}.\end{aligned}$$

The lower bound becomes tight when the instrumental distribution is the posterior distribution

$$q(\mathbf{S}) = p(\mathbf{S}|\mathbf{X}, \mathbf{T}, \mathbf{V})$$

which for posterior distributions in the exponential family, can be calculated and represented solely in terms of its sufficient statistics. We can thus maximize the log likelihood iteratively by means of coordinate ascent. Calculating the instrumental distribution by means of its sufficient statistics is known as the expectation step, and then the maximization step refers to the maximization of the bound by coordinate ascent. The EM algorithm can be expressed as follows at iteration i: the expectation step is

$$q\left(\mathbf{S}^{(i)}|\,\{\mathbf{T}, \mathbf{V}\}^{(i-1)}\right) = p\left(\mathbf{S}|\mathbf{X}, \{\mathbf{T}, \mathbf{V}\}^{(i-1)}\right) \quad (3)$$

and the maximization step is

$$\{\mathbf{T}, \mathbf{V}\}^{(i)} = \arg \max_{\{\mathbf{T}, \mathbf{V}\}} \left\langle \log p(\mathbf{X}, \mathbf{S}, \mathbf{T}, \mathbf{V}) \right\rangle. \quad (4)$$

The expectation in (4) is with respect to the instrumental distribution $q\left(\mathbf{S}^{(i)}|\,\{\mathbf{T}, \mathbf{V}\}^{(i-1)}\right)$ calculated in the previous expectation step (3). The maximization step for these matrix factorization models typically cannot be computed in a single step. Iterative solutions are required, and it has been shown by Neal and Hinton [14] that replacing the maximization step with a step that merely increases the likelihood, rather than maximizing the likelihood, is sufficient for convergence. In the following sections, we will describe two probabilistic source models with conjugate priors in the exponential family. We derive the conditional posterior distributions of the $\mathbf{T}$ and $\mathbf{V}$ parameters, and are able to then increase the log likelihood in (4) by updating these parameters to be equal to the modes of the conditional posterior distributions. We are permitted to perform as many updates of $\mathbf{T}$ and $\mathbf{V}$ within the maximization step as desired, before computing the expectations of the sources again, provided we confirm that the log likelihood has indeed increased with each iteration.

The EM algorithm is partly Bayesian in that it maintains distributions over the sources and point estimates over the parameters. We can instead adopt a fully Bayesian approach, which additionally maintains distributions over the parameters, by approximating the posterior distribution with a factored instrumental distribution $q(\mathbf{S}, \mathbf{T}, \mathbf{V}) = q(\mathbf{S})\, q(\mathbf{T}), q(\mathbf{V})$. This type of approximation is known as the mean-field approximation or Variational Bayes (VB) [15]. In terms of implementing the algorithm, we calculate the sufficient statistics of the parameters in the VB method, rather than calculating the mode in the EM method.

C. Gaussian Variance Model

This model assumes that each element of the observation matrix $\mathbf{X}$ is distributed zero-mean normal, with the variance given by the elements of $\mathbf{TV}$. This source model has been applied in audio source separation in [29] and time–frequency estimation previously in [16]. An EM algorithm for the Gaussian variance model has been presented in [30]. The likelihood is

$$p\left(\mathbf{X}|\mathbf{TV}\right) = \prod_{\nu,k} \mathcal{N}\left(\mathbf{X}_{\nu,k}; 0, \left[\mathbf{TV}\right]_{\nu,k}\right).$$

As the elements of the template and excitation matrices are used as the variance parameters of a normal distribution, we find it convenient to represent prior information concerning these parameters using inverse-gamma distributions, which is the conjugate prior for the variance of a normal distribution. We use $\mathcal{IG}(r, \alpha, \beta)$ to denote that r has an inverse-gamma distribution with shape α and scale β. The priors we use for the template and excitation matrices are

$$p(\mathbf{T}) = \prod_{\nu,n} \mathcal{IG}\left(\mathbf{T}_{\nu,n}; \alpha^{(\mathbf{T})}_{\nu,n}, \beta^{(\mathbf{T})}_{\nu,n}\right)$$
$$p(\mathbf{V}) = \prod_{n,k} \mathcal{IG}\left(\mathbf{V}_{n,k}; \alpha^{(\mathbf{V})}_{n,k}, \beta^{(\mathbf{V})}_{n,k}\right).$$

To derive the expectation step, we require the conditional distribution of the sources given the parameters. The posterior distribution of the sources can be factorized into independent distributions over the vector of source coefficients for each individual time–frequency bin:

$$q\left(\mathbf{S}|\mathbf{T},\mathbf{V}\right) = p\left(\mathbf{S}|\mathbf{X},\mathbf{T},\mathbf{V}\right) = \prod_{\nu,k} p\left(\mathbf{s}_{\nu,k}|\mathbf{X}_{\nu,k}, \mathbf{t}_\nu \mathbf{v}_k^\top\right) \quad (5)$$

where the nth element of the vector $\mathbf{s}_{\nu,k}$ is $\left[\mathbf{S}_n\right]_{\nu,k}$, $\mathbf{t}_\nu$ is the νth column vector of $\mathbf{T}$, and $\mathbf{v}_k$ is the kth row vector of $\mathbf{V}$. Note that $\left[\mathbf{TV}\right]_{\nu,k} = \text{Trace}\left[\mathbf{t}_\nu \mathbf{v}_k^\top\right]$. Each vector has a multivariate normal distribution, for which the sufficient statistics can be expressed compactly. Define the vector of responsibilities as

$$\kappa_{\nu,k} = \frac{1}{\left[\mathbf{TV}\right]_{\nu,k}} \begin{bmatrix} \left[\mathbf{t}_\nu \mathbf{v}_k^\top\right]_{1,1} \\ \vdots \\ \left[\mathbf{t}_\nu \mathbf{v}_k^\top\right]_{N,N} \end{bmatrix} \quad (6)$$

then the mean value of $\mathbf{s}_{\nu,k}$ under (5) is simply the observation weighted by the responsibilities

$$\left\langle \mathbf{s}_{\nu,k} \right\rangle_{q(\mathbf{S}|\mathbf{T},\mathbf{V})} = \kappa_{\nu,k} \mathbf{X}_{\nu,k}$$

and the correlation matrix of $\mathbf{s}_{\nu,k}$ under (5) is

$$\left\langle \mathbf{s}_{\nu,k}\mathbf{s}_{\nu,k}^\top \right\rangle_{q(\mathbf{S}|\mathbf{T},\mathbf{V})} = \left[\mathbf{t}_\nu \mathbf{v}_k^\top\right] \cdot I_N - \kappa_{\nu,k}\kappa_{\nu,k}^\top \left[\mathbf{TV}\right]_{\nu,k} + \left\langle \mathbf{s}_{\nu,k} \right\rangle \left\langle \mathbf{s}_{\nu,k} \right\rangle^\top.$$

The maximization rules are most conveniently derived by considering the conditional distributions of the posterior. As the priors are conjugate, these conditional distributions are themselves inverse-gamma. Collecting the terms of the joint distribution dependent on the templates we have

$$\begin{aligned} &\left\langle \log p\left(\mathbf{X},\mathbf{S},\mathbf{T},\mathbf{V}\right) \right\rangle \\ &\quad = -\sum_{\nu,n} \log \mathbf{T}_{\nu,n} (\alpha^{(\mathbf{T})}_{\nu,n} + K + 1) \\ &\qquad - \sum_{\nu,n} \mathbf{T}^{-1}_{\nu,n} \left(\beta^{(\mathbf{T})}_{\nu,n} + \sum_k \mathbf{V}^{-1}_{n,k} \left\langle \mathbf{s}_{\nu,k}\mathbf{s}_{\nu,k}^\top \right\rangle_{n,n}\right) + \cdots \end{aligned}$$

and collecting the excitation terms we have

$$\begin{aligned} &\left\langle \log p(\mathbf{X},\mathbf{S},\mathbf{T},\mathbf{V}) \right\rangle \\ &\quad = -\sum_{n,k} \log \mathbf{V}_{n,k} (\alpha^{(\mathbf{V})}_{n,k} + F + 1) \\ &\qquad - \sum_{n,k} \mathbf{V}^{-1}_{n,k} \left(\beta^{(\mathbf{V})}_{n,k} + \sum_\nu \mathbf{T}^{-1}_{\nu,n} \left\langle \mathbf{s}_{\nu,k}\mathbf{s}_{\nu,k}^\top \right\rangle_{n,n}\right) + \cdots \end{aligned}$$

where the expectations in the above expressions are with respect to the instrumental distribution $q\left(\mathbf{S}|\left\{\mathbf{T},\mathbf{V}\right\}^{(i-1)}\right)$. It can be seen that these expectations are inverse-gamma distributions, and we can update the parameters to be equal to the mode of these distributions. A full algorithmic description is provided in Fig. 1.

- **Source Expectation**

  $\left\langle \mathbf{s}_{\nu,k}\mathbf{s}_{\nu,k}^\top \right\rangle = \left[\mathbf{t}_\nu \mathbf{v}_k^\top\right] \cdot I_N - \kappa_{\nu,k}\kappa_{\nu,k}^\top \left[\mathbf{TV}\right]_{\nu,k} + \left\langle \mathbf{s}_{\nu,k} \right\rangle \left\langle \mathbf{s}_{\nu,k} \right\rangle^\top$

- **Template Maximization**
  Shape and scale parameters of inverse-gamma posterior distribution

  $$\begin{aligned} A_{\nu,n} &= \alpha^{(\mathbf{T})}_{\nu,n} + K \\ B_{\nu,n} &= \beta^{(\mathbf{T})}_{\nu,n} + \sum_k \mathbf{V}^{-1}_{n,k} \left\langle \mathbf{s}_{\nu,k}\mathbf{s}_{\nu,k}^\top \right\rangle \end{aligned}$$

  Mode of posterior distribution

  $$\mathbf{T}_{\nu,n} \leftarrow \frac{B_{\nu,n}}{A_{\nu,n}+1}$$

- **Excitation Maximization**
  Shape and scale parameters of inverse-gamma posterior distribution

  $$\begin{aligned} A_{n,v} &= \sum_{\{k:\mathbf{C}_{nk}=v\}} \alpha^{(\mathbf{V})}_{n,k} + F|\{k:\mathbf{C}_{nk}=v\}| \\ B_{n,v} &= \sum_{\{k:\mathbf{C}_{nk}=v\}} \left(\beta^{(\mathbf{V})}_{n,k} + \sum_\nu \mathbf{T}^{-1}_{\nu,n} \left\langle \mathbf{s}_{\nu,k}\mathbf{s}_{\nu,k}^\top \right\rangle\right) \end{aligned}$$

  Mode of posterior distribution

  $$\mathbf{F}_{n,v} \leftarrow \frac{B_{n,v}}{A_{n,v}+1}$$

- **Transcription Search**
  for $k = 1, \ldots, K$

  $$\mathbf{C}_k \leftarrow \arg\max_{\tilde{\mathbf{C}}_k} \sum_\nu \left(-\frac{1}{2}\frac{|\mathbf{X}_{\nu,k}|^2}{[\mathbf{T}\tilde{\mathbf{V}}]_{\nu,k}} - \log[\mathbf{T}\tilde{\mathbf{V}}]_{\nu,k}\right) p(\mathbf{F}, \tilde{\mathbf{C}}_k)$$

Fig. 1. Gaussian Variance: algorithm for polyphonic transcription.

D. Poisson Intensity Model

In the previous section, we placed a variance parameter on the coefficients of the spectrogram. The Poisson model on the

other hand assigns an intensity parameter to the (non-negative) magnitude of the spectrogram. This is a probabilistic interpretation of the divergence measure

$$D(\mathbf{X}||\mathbf{TV}) = -\sum_{\nu,k}\left(\mathbf{X}_{\nu,k}\log\frac{[\mathbf{TV}]_{\nu,k}}{\mathbf{X}_{\nu,k}} - [\mathbf{TV}]_{\nu,k} + \mathbf{X}_{\nu,k}\right)$$

where in this section we use $\mathbf{X}$ to refer to the squared magnitude of the spectrogram coefficients $\left|x_{\nu,k}^2\right|$. This measure has be shown by Smargadis and Brown [10] to have better properties for music transcription in appropriately distributing the energy of the signal into the correct sources than when using the Frobenius norm measure. A simple algorithm involving iterative application of matrix update rules has been described in [17] to minimize this divergence measure, and this has been shown in [11], [18], [19] to be equivalent to the EM algorithm for maximizing the likelihood, as mentioned in the original NMF papers [17], [20]

$$p\left(\mathbf{X};\mathbf{TV}\right) = \prod_{\nu,k}\mathcal{P}o\left(\mathbf{X}_{\nu,k};[\mathbf{TV}]_{\nu,k}\right)$$

where $\mathcal{P}o(r;\lambda)$ denotes that r has a Poisson distribution with intensity λ. In order to satisfy (1), it can be verified that $\mathbf{S}_n \sim \mathcal{P}o(\mathbf{S}_n;\mathbf{t}_n\mathbf{v}_n^\top)$.

In an analogous manner to the variance model, we can put gamma prior distributions on the parameters in a Bayesian setting. We use $\mathcal{G}(r;\alpha,\beta)$ to denote that r has a gamma distribution with shape α and rate β. The priors we use for the template and excitation matrices are

$$p(\mathbf{T}) = \prod_{\nu,n}\mathcal{G}(\mathbf{T}_{\nu,n};\alpha_{\nu,n}^{(\mathbf{T})},\beta_{\nu,n}^{(\mathbf{T})})$$
$$p(\mathbf{V}) = \prod_{n,k}\mathcal{G}(\mathbf{V}_{n,k};\alpha_{n,k}^{(\mathbf{V})},\beta_{n,k}^{(\mathbf{V})}).$$

To derive the expectation rule, we use the result that the posterior distribution of the latent sources is multinomial [19], and the mean value is again the observation weighted by the responsibilities

$$\left\langle\mathbf{s}_{\nu,k}\right\rangle_{q(\mathbf{S}|\mathbf{T},\mathbf{V})} = \kappa_{\nu,k}\mathbf{X}_{\nu,k}$$

where the responsibilities are defined the same as for the Gaussian model (6). This particular result highlights both the similarity in the construction of the variance and intensity models, but also a weakness in the generative model with the Poisson assumption. Both models construct the sources by weighting the observations according to their relative energy, however the variance model weights the coefficients themselves, which means the sources themselves have a physical interpretation, while the intensity model weights the magnitude of the coefficients, which is not physically realistic as the magnitudes of the sources do not superimpose in practice to result in the observations. Hence, the variance model is able to model effects such as cancellation.

- **Source Expectation**

$$\left\langle\mathbf{s}_{\nu,k}\right\rangle = \kappa_{\nu,k}\mathbf{X}_{\nu,k}$$

- **Template Maximization**
Shape and scale parameters of inverse-gamma posterior distribution

$$A_{\nu,n} = \alpha_{\nu,n}^{(\mathbf{T})} + \sum_k\left\langle\mathbf{s}_{\nu,k}\right\rangle$$
$$B_{\nu,n} = \beta_{\nu,n}^{(\mathbf{T})} + \sum_k\mathbf{V}_{n,k}$$

Mode of posterior distribution

$$\mathbf{T}_{\nu,n} \leftarrow \frac{A_{\nu,n}-1}{B_{\nu,n}}$$

- **Excitation Maximization**
Shape and scale parameters of inverse-gamma posterior distribution

$$A_{n,v} = \sum_{\{k:\mathbf{C}_{nk}=v\}}\left(\alpha_{n,k}^{(\mathbf{V})} + \sum_\nu\left\langle\mathbf{s}_{\nu,k}\right\rangle\right)$$
$$B_{n,v} = \sum_{\{k:\mathbf{C}_{nk}=v\}}\left(\beta_{n,k}^{(\mathbf{V})} + \sum_\nu\mathbf{T}_{\nu,n}\right)$$

Mode of posterior distribution

$$\mathbf{F}_{n,v} \leftarrow \frac{A_{n,v}-1}{B_{n,v}}$$

- **Transcription Search**
for $k = 1,\ldots,K$

$$\mathbf{C}_k \leftarrow \arg\max_{\tilde{\mathbf{C}}_k}\sum_\nu\left(\mathbf{X}_{\nu,k}\log[\mathbf{T}\tilde{\mathbf{V}}]_{\nu,k} - [\mathbf{T}\tilde{\mathbf{V}}]_{\nu,k}\right)p(\mathbf{F},\tilde{\mathbf{C}}_k)$$

Fig. 2. Poisson Intensity: algorithm for polyphonic transcription.

The maximization rules result again from the conditional distributions of the posterior, which are gamma. Collecting the terms for the templates in the joint distribution, we have

$$\begin{aligned}&\left\langle\log p(\mathbf{X},\mathbf{S},\mathbf{T},\mathbf{V})\right\rangle\\ &\quad= \sum_{\nu,n}\log\mathbf{T}_{\nu,n}(\alpha_{\nu,n}^{(\mathbf{T})} + \sum_k\left\langle\mathbf{s}_{\nu,k}\right\rangle - 1)\\ &\qquad- \sum_{\nu,n}\mathbf{T}_{\nu,n}^{-1}\left(\beta_{\nu,n}^{(\mathbf{T})} + \sum_k\mathbf{V}_{n,k}\right) + \cdots\end{aligned}$$

and collecting the excitation terms we have

$$\begin{aligned}&\left\langle\log p(\mathbf{X},\mathbf{S},\mathbf{T},\mathbf{V})\right\rangle\\ &\quad= \sum_{n,k}\log\mathbf{V}_{n,k}(\alpha_{n,k}^{(\mathbf{V})} + \sum_\nu\left\langle\mathbf{s}_{\nu,k}\right\rangle - 1)\\ &\qquad- \sum_{n,k}\mathbf{V}_{n,k}^{-1}\left(\beta_{n,k}^{(\mathbf{V})} + \sum_\nu\mathbf{T}_{\nu,n}\right) + \cdots\end{aligned}$$

where the expectations are with respect to $q\left(\mathbf{S}|\left\{\mathbf{T},\mathbf{V}\right\}^{(i-1)}\right)$. These expectations are gamma distributions, and we can update the parameters to be equal to the mode of these distributions. A full algorithmic description is provided in Fig. 2.

III. Prior Model for Polyphonic Piano Music

In this section, we extend the prior model for the excitation matrix to include MIDI pitch and velocity of the notes that are playing in a piece of solo polyphonic piano music.

A. Model Description

In this paper, we have chosen to rely on deterministic approaches to solve the transcription inference problem, as opposed to more expensive Monte Carlo approaches [21]. In this section, we describe a quite general approach which lends itself to any form of music for which the MIDI format is an admissible representation of the transcription.

We select the maximum number of sources N to be the total number of pitches represented in the MIDI format. Each source n corresponds to a particular pitch. Then we have a single set of template parameters $\mathbf{T} \in \mathbb{R}_+^{F\times N}$ for all n sources, which are intended to represent the spectral, harmonic information of the pitches. For polyphonic transcription, we are typically interested in inferring the piano roll matrix $\mathbf{C}$ which owing to the above assumption of one source per pitch has the same dimensions as the excitation matrix $\mathbf{V}$. For note n at time k we set $\mathbf{C}_{n,k}$ to be the value of the velocity of the note, and $\mathbf{C}_{n,k} = 0$ if the note is not playing. We use the NOTE ON velocity, which is stored in the MIDI format as a integer between 1 and 128. Thus, we model note velocity using our generative model. This contrasts with previous approaches which infer a binary-valued piano roll matrix of note activity, essentially discarding potentially useful volume information. The prior distribution $p(\mathbf{C})$ is a discrete distribution, which can incorporate note transition probabilities and commonly occurring groups of note pitches, i.e., chords and harmony information.

Our intuition is that a note with a larger velocity will have a larger corresponding excitation. The magnitude of the excitation will depend on the pitch of the note as well as its velocity. We will represent this information as a set of *a priori* unknown positive-valued random vectors $\mathbf{f}_n \in \mathbb{R}_+^{128}$. In words, the values of $\mathbf{f}_n$ represent a mapping from the MIDI pitch and velocity to the excitation matrix. For music transcription, we extend the prior model on $\mathbf{V}$ to include $\mathbf{F} = \{\mathbf{f}_n\}, n = 1, \ldots, N$ and $\mathbf{C}$. We have

$$p(\mathbf{V}, \mathbf{F}, \mathbf{C}) = p(\mathbf{V}|\mathbf{F}, \mathbf{C})p(\mathbf{F}, \mathbf{C})$$

and the mapping itself is given by

$$\mathbf{V} = \begin{cases} 0, & \mathbf{C}_{nk} = 0 \\ \mathbf{f}_n[\mathbf{C}_{nk}], & \text{otherwise.} \end{cases}$$

As $\mathbf{F}$ is a mapping to the excitation matrix, we place an inverse-gamma prior (for the Gaussian variance model) or a gamma prior (for the Poisson intensity model) over each element of $\mathbf{F}$. The resulting conditional posterior over $\mathbf{F}$ is of the same family as the prior, and is obtained by combining the expectations of the sources corresponding to the correct pitch and velocity.

The full generative model for polyphonic transcription is given by

$$p(\mathbf{X}, \mathbf{S}, \mathbf{T}, \mathbf{V}, \mathbf{F}, \mathbf{C}) = p(\mathbf{X}|\mathbf{S})p(\mathbf{S}|\mathbf{T}, \mathbf{V})p(\mathbf{V}|\mathbf{F}, \mathbf{C})p(\mathbf{T})p(\mathbf{F}, \mathbf{C}).$$

One advantage of this model is that minimal storage is required for the parameters which can be estimated offline from training data, as we demonstrate in Section IV-B. The two sets of parameters are intuitive for musical signals. This model also allows closer modeling of the excitation of the notes that the MIDI format allows.

B. Algorithm

The algorithm we use is a generalized EM algorithm. We iterate towards the maximum *a posteriori* solution of the marginal likelihood

$$\arg\max_{\mathbf{T},\mathbf{F},\mathbf{C}} p(\mathbf{X}, \mathbf{T}, \mathbf{F}, \mathbf{C})$$

by marginalizing the latent parameters $\mathbf{S}$ which has been covered in Section II-B, and $\mathbf{V}$ which is straightforward given that $p(\mathbf{V}|\mathbf{F}, \mathbf{C})$ is deterministic. The posterior distribution of $\mathbf{F}$ is inverse-gamma as it is formed by collecting the estimates of $\mathbf{V}$ corresponding to each note pitch/velocity pairing.

To maximize for the piano roll $\mathbf{C}$ we first note that each frame of observation data is independent given the other parameters $\mathbf{V}, \mathbf{F}$. For each k we wish to calculate

$$\arg\max_{\mathbf{C}_k} p(\mathbf{X}_k|\mathbf{T}, \mathbf{V}_k)p(\mathbf{V}_k|\mathbf{F}, \mathbf{C}_k)p(\mathbf{F}, \mathbf{C}_k).$$

However, as each $\mathbf{C}_k$ has N^{128} possible values, an exhaustive search to maximize this is not feasible. Instead, we have found that the following greedy search algorithm works sufficiently well: for each frame k calculate

$$\arg\max_{\tilde{\mathbf{C}}_k} p(\mathbf{X}_k|\mathbf{T}, \tilde{\mathbf{V}}_k)p(\tilde{\mathbf{V}}_k|\mathbf{F}, \tilde{\mathbf{C}}_k)p(\mathbf{F}, \tilde{\mathbf{C}}_k)$$

where $\tilde{\mathbf{C}}_k$ differs from $\mathbf{C}_k$ by at most one element, and $\tilde{\mathbf{V}}$ is the corresponding excitation matrix. There are $N\times 128$ possible settings of $\tilde{\mathbf{C}}_k$ for which we evaluate the likelihood at each stage of the greedy search. This can be carried out efficiently by noticing that during the search the corresponding matrix products $\mathbf{T}\tilde{\mathbf{V}}$ differ from the existing value by only a rank-one update of $\mathbf{TV}$.

The resulting algorithm has one update for the expectation step and three possible updates for the maximization step. For the generalized EM algorithm to be valid, we must ensure that any maximization step based on parameter values not used to calculate the source expectations is not guaranteed to increase the log likelihood, and therefore must be verified.

IV. Results

A. Comparison

To comprehensively evaluate these models, we use Poliner and Ellis training and test data [1] and compare the performance against the results provided in the same paper, which are repeated here for convenience. The ground truth for the data consists of 124 MIDI files of classical piano music, of which 24 have been designated for testing purposes and 13 are designated for validation. In a Bayesian framework there is no distinction between training and validation data: both are considered labeled observations. Here we have chosen to discard the validation data rather than include it in the training examples for a fairer comparison with the approaches used by other authors. Only the first 60 s of each extract is used.

The observation data is primarily obtained by using a software synthesizer to generate audio data. In addition, 19 of the training tracks and ten of the test tracks were synthesized

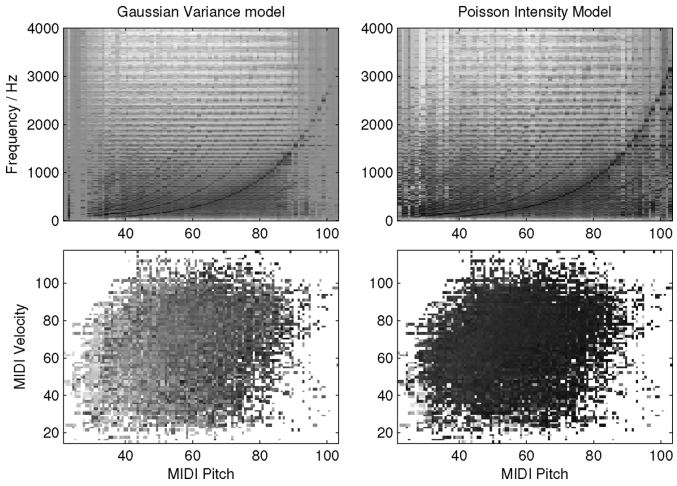

Fig. 3. $\log \mathbf{T}$ templates (upper row) and $\log \mathbf{F}$ excitation (lower row) parameters estimated from training data for the Gaussian variance and Poisson intensity models, with flat prior distributions. Both models capture the harmonicity present in musical pitches in the spectral templates, and the excitation mapping increases with increasing note velocity. For the excitation parameters, white areas denote pitch/velocity pairs that are not present in the training data and are thus unobserved.

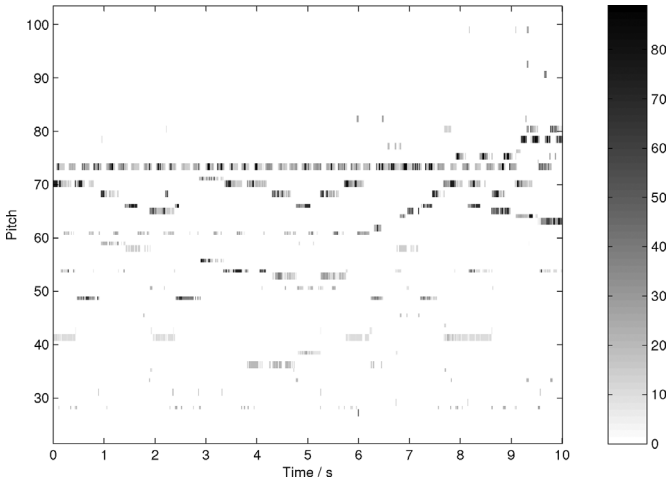

Fig. 4. Transcription with independent prior on $\mathbf{C}$. The generative model has not only detected the activity of many of the notes playing, but also has attempted to jointly infer the velocity of the notes. Each frame has independently inferred velocity, hence there is much variation across a note, however there is correlation between the maximum inferred velocity during a note event and the ground truth velocities.

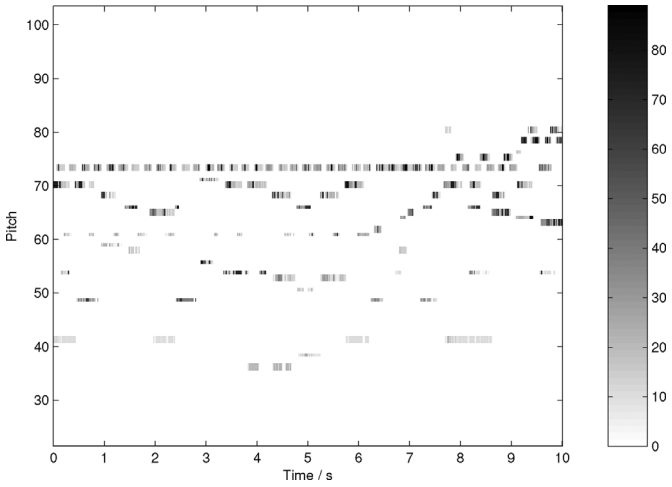

Fig. 5. Transcription with Markov prior on $\mathbf{C}$. The Markov prior on $\mathbf{C}$ has eliminated many of the spurious notes detected, which are typically of a short duration of a few frames.

and recorded on a Yamaha Disklavier. The audio, sampled at 8000 Hz, is then buffered into frames of length 128 ms with a 10 ms hop between frames, and the spectrogram is obtained from the short-time Fourier transform of these frames. Poliner and Ellis subsequently carry out a spectral normalization step in order to remove some of the timbral and dynamical variation in the data prior to classification. However, we omit this processing stage as we rather wish to capture this information in our generative model.

B. Implementation

Because of the copious amount of training data available, there is enough information concerning the frequencies of the occurrence of the note pitches and velocities that it is not necessary to place informative priors on these parameters.

It is not necessary to explicitly carry out a training run to estimate values of the model parameters before evaluating against the test data. However, the EM algorithm does converge faster during testing if we first estimate the parameters from the training data. Fig. 3 shows the parameters under the two models after running the EM algorithm to convergence on the training data only. The templates clearly exhibit the harmonic series of the musical notes, and the excitations contain the desired property that notes with higher velocity correspond to higher excitation, hence our assumption of flat priors on these parameters seems appropriate.

For each of the matrix factorization models we consider two choices of the prior $\mathbf{C}$. The first assumes that each frame of data is independent of the others, which is useful in evaluating the performance of the source models in isolation. The second assumes that each note pitch is independent of the others, and between consecutive frames there is a state transition probability, where the states are each note being active or inactive, i.e.,

$$p(\mathbf{C}_{n,k}>0|\mathbf{C}_{n,k-1}=0)=p(\mathbf{C}_{n,k}=0|\mathbf{C}_{n,k-1}>0)=p_{\text{event}}.$$

The state transition probabilities are estimated from the training data. It is possible and more correct to include these transition probabilities as parameters in the model, but we have not carried out the inference of note transition probabilities in this work.

C. Evaluation

Following training, the matrix of spectrogram coefficients is then extended to include the test extracts. As the same two instruments are used in the training and test data, we simply use the same parameters which were estimated in the training phase. We transcribe each test extract independently of the others, yet note that in the full Bayesian setting this should be carried out jointly; however, this is not practical or typical of a reasonable application of a transcription system. An example of the transcription output for the first ten seconds of the synthesized version of Burgmueller's *The Fountain* is provided for the Gaussian

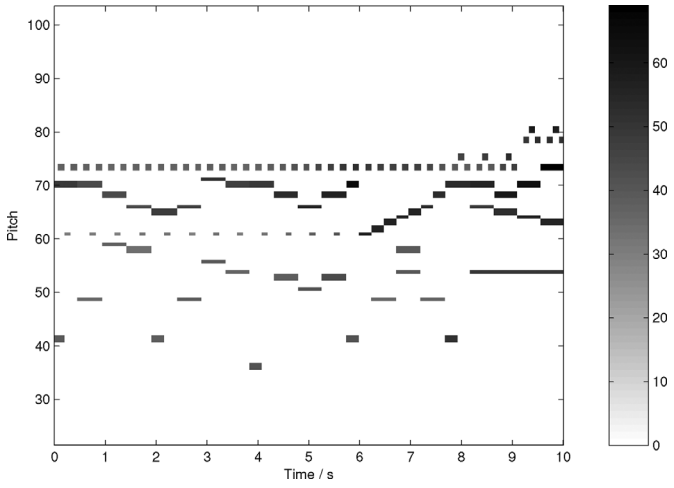

Fig. 6. Ground truth for the extract transcribed in Figs. 4 and 5. We have used only the information contained in note pitches, but the effect of resonance and pedaling can be clearly seen in the transcriptions. This motivates the use of a note onset evaluation criteria.

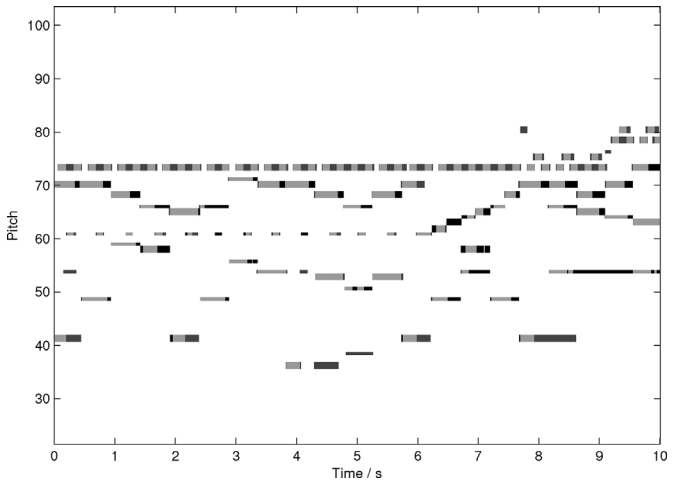

Fig. 7. Detection assessment. True positives are in light gray, false positives in dark gray, and false negatives in black. Most of the difficulties encountered in transcription in this particular extract were due to the positioning of note onsets and offsets, rather than the detection of the pitches themselves.

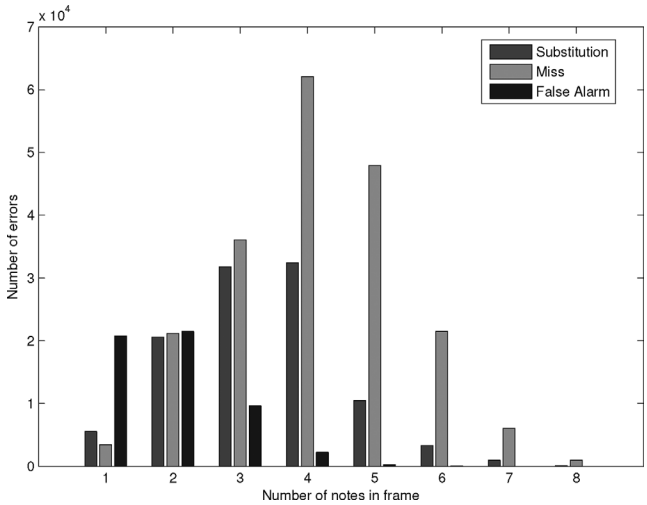

Fig. 8. Number of errors for the Gaussian variance Markov model, categorized by number of notes in a frame and by error type.

TABLE I
FRAME-LEVEL TRANSCRIPTION ACCURACY

Model	Piano	MIDI	Both
SVM	56.5	72.1	67.7
Ryynänen & Klapuri	41.2	48.3	46.3
Marolt	38.4	40.0	39.6
Variance (Independent)	36.0	41.2	39.7
Variance (Markov)	38.0	44.0	42.3
Intensity (Independent)	40.1	35.4	36.8
Intensity (Markov)	39.7	36.2	37.3

TABLE II
FRAME-LEVEL TRANSCRIPTION RESULTS

Model	ACC	TOT	SUBS	MISS	FA
SVM	67.7	34.2	5.3	12.1	16.8
Ryynänen & Klapuri	46.6	52.3	15.0	26.2	11.1
Marolt	36.9	65.7	19.3	30.9	15.4
Variance (Independent)	39.7	68.2	22.9	27.7	17.6
Variance (Markov)	42.3	62.1	18.1	32.0	12.0
Intensity (Independent)	36.8	71.0	27.8	24.6	18.6
Intensity (Markov)	37.3	66.6	23.7	30.0	12.9

variance model, both with independent (Fig. 4) and Markov priors (Fig. 5) on **C**, compared to the MIDI ground truth (Fig. 6). The transcription is graphically represented in terms of detections and misses in Fig. 7. We follow the same evaluation criteria as provided by Poliner and Ellis. As well as recording the accuracy ACC (true positive rate), the transcription is error is decomposed into three parts: SUBS the substitution error rate, when a note from the ground truth is transcribed with the wrong pitch; MISS the note miss rate, when a note in the ground truth is not transcribed, and FA the false alarm rate beyond substitutions, when a note not present in the ground truth is transcribed. These sum to form the total transcription error TOT which cannot be biased simply by adjusting a threshold for how many notes are transcribed.

Table I shows the frame-level transcription accuracy for the approaches studied in [1]. We are using the same data sets and features dimensions selected by the authors of this paper to compare our generative models against these techniques. This table expands the accuracy column in Table II by splitting the test data into the recorded piano extracts and the MIDI synthesized extracts.

Table II shows the frame-level transcription results for the full synthesized and recorded data set. Accuracy is the true positive rate expressed as a percentage, which can be biased by not reporting notes. The total error is a more meaningful measure which is divided between substitution, note misses, and false alarm errors. This table shows that the matrix factorization models with a Markov note event prior have a similar error rate to the Marolt system on this dataset, but has a greater error rate than the support vector machine classifier. Fig. 8 shows how the error varies with different numbers of notes in a frame.

V. CONCLUSION AND FURTHER IMPROVEMENTS

We have compared the performance of generative spectrogram factorization models with three existing transcription systems on a common dataset. The models exhibit a similar error rate as the neural-network classification system of [3]. However,

the support vector machine classifier of [1] achieves a lower error rate for polyphonic piano transcription on this dataset. In this conclusion, we principally discuss the reasons for the difference in error rate of these systems, and how the generative models can be improved in terms of inference and prior structure to achieve an improved performance.

The support vector machine is purely a classification system for transcription, for which the parameters have been explicitly chosen to provide the best transcription performance on a validation set; while the spectrogram factorization models, being generative in nature, are applicable to a much wider range of problems: source separation, restoration, score-audio alignment, and so on. For this reason, we have not attempted to select priors by hand-tuning in order to improve transcription performance, but rather adopt a fully Bayesian approach with an explicit model which infers correlations in the spectrogram coefficients in training and test data, and thus as a product of this inference provides a transcription of the test data. The differences in this style of approach, and the subsequent difference in performance, resemble that of supervised and unsupervised learning in classification. Thus, in light of this, we consider the performance of the spectrogram factorization models to be encouraging, as they are comparable to an existing polyphonic piano transcription system without explicitly attempting to improve the transcription performance by tuning prior hyperparameters. Vincent *et al.* [22], for instance, demonstrate the improvement in performance for polyphonic piano transcription that can be achieved over the standard NMF algorithm by developing improved basis spectra for the pitches, and achieve a performance mildly better than the neural-network classifier: a similar result to what has been presented here, and conclude that an NMF-based system is competitive in the MIREX classification task.

To improve performance for transcription in a Bayesian spectrogram factorization, we can first improve initialization using existing multiple frequency detection systems for spectrogram data, and extend the hierarchical model for polyphonic transcription using concepts such as chords, keys. We can also jointly track tempo and rhythm using a probabilistic model; for examples of this see [23]–[25], where the model used could easily be incorporated into the Bayesian hierarchical approach here.

The models we have used have assumed that the templates and excitations are drawn independently from priors; however, the existing framework of gamma Markov fields [26]–[28] can be used as replacements of these priors, and allows us to model stronger correlations, for example, between the harmonic frequencies of the same musical pitch, which additionally contain timbral content, and also model the damping of the excitation of notes from one frame to the next. It has qualitatively shown that using gamma Markov field priors results in a much improved transcription, and in future work we will use this existing framework to extend the model described in this paper, expecting to see a much improved transcription performance by virtue of a more appropriate model of the time–frequency surface.

On this dataset, the Gaussian variance model has better performance for transcription than the intensity-based model, and we suggest that this is due to the generative model modeling the weighting of the spectrogram coefficients directly, and thus being a more appropriate model for time–frequency surface estimation. However, most of the literature for polyphonic music transcription systems using matrix factorization models has focused on the KL divergence and modifications and enhancements of the basic concept. Therefore, it would be useful to first evaluate such variants of NMF against this dataset and other systems used for comparing and evaluating music transcription systems. Second, it would also be useful to replace the implicit Poisson intensity source model in these approaches with the Gaussian variance model, to the advantage of the better generative model.

In this paper, we have derived a generalized expectation-maximization algorithm for generative spectrogram factorization models. However, with such schemes we experience slow convergence to local maxima. Performance can be improved using Monte Carlo methods [21] to generate samples from the posterior distribution, using proposal distributions designed from multiple frequency detector algorithms. Furthermore, inference can be performed in an online manner for applications that require this.

In summary, we have presented matrix factorization models for spectrogram coefficients, using Gaussian variance and Poisson intensity parametrization, and have developed inference algorithms for the parameters of these models. The suitability of these models has been assessed for the polyphonic transcription of solo piano music, resulting in a performance which is comparable to some existing transcription systems. As we have used a Bayesian approach, we can extend the prior structure in a hierarchical manner to improve performance and model higher-level features of music.

REFERENCES

[1] G. E. Poliner and D. P. W. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP J. Adv. Signal Process.*, vol. 2007, pp. 154–162, 2007.

[2] G. E. Poliner and D. P. W. Ellis, "Improving generalization for classification-based polyphonic piano transcription," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, 2007, pp. 86–89.

[3] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Trans. Multimedia*, vol. 6, no. 3, pp. 439–449, Jun. 2004.

[4] M. P. Ryynänen and A. Klapuri, "Polyphonic music transcription using note event modeling," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, 2005, pp. 319–322.

[5] A. T. Cemgil, B. Kappen, and D. Barber, "Generative model based polyphonic music transcription," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, 2003, pp. 181–184.

[6] A. T. Cemgil, B. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 679–694, Mar. 2006.

[7] K. Kashino and S. J. Godsill, "Bayesian estimation of simultaneous musical notes based on frequency domain modelling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montreal, QC, Canada, 2004, pp. IV–305–IV–308.

[8] S. J. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, 2002, pp. 1769–1772.

[9] S. A. Abdallah and M. D. Plumbley, "Unsupervised analysis of polyphonic music by sparse coding," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 179–196, Jan. 2006.

[10] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, 2003, pp. 177–180.

[11] T. Virtanen, A. T. Cemgil, and S. J. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, Las Vegas, NV, 2008, pp. 1825–1828.
[12] E. Vincent and M. D. Plumbley, "Efficient Bayesian inference for harmonic models via adaptive posterior factorization," *Neurocomputing*, vol. 72, pp. 79–87, Dec. 2008.
[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist., Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.
[14] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*. Cambridge, MA, USA: MIT Press, 1999, pp. 355–368.
[15] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, pp. 183–233, Nov. 1999.
[16] P. Wolfe, S. J. Godsill, and W. Ng, "Bayesian variable selection and regularization for time–frequency surface estimation," *J. R. Statist. Soc. Series B*, vol. 66, pp. 575–589, Aug. 2004.
[17] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Adv. Neural Inf. Process. Syst.*, pp. 556–562, 2001.
[18] H. Kameoka, "Statistical approach to multipitch analysis," Ph.D., Univ. Tokyo, Tokyo, Japan, 2007.
[19] A. T. Cemgil, "Bayesian inference in non-negative matrix factorisation models," Dept. Eng., Univ. Cambridge, U.K., 2008, Tech. Rep. CUED/F-INFENG/TR.609.
[20] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.
[21] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 1999.
[22] E. Vincent, N. Berlin, and R. Badeau, "Harmonic and inharmonic non-negative matrix factorization for polyphonic pitch transcription," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, 2008, pp. 109–112.
[23] N. Whiteley, A. T. Cemgil, and S. J. Godsill, "Bayesian modelling of temporal structure in musical audio," in *Proc. 7th Int. Conf. Music Inf. Retrieval*, Victoria, BC, Canada, 2006, pp. 29–34.
[24] C. Raphael, "A hybrid graphical model for aligning polyphonic audio with musical scores," in *Proc. 5th Int. Conf. Musical Inf. Retrieval*, Barcelona, Spain, 2004, pp. 387–394.
[25] P. Peeling, A. T. Cemgil, and S. J. Godsill, "A probabilistic framework for matching music representations," in *Proc. 8th Int. Conf. Music Inf. Retrieval*, Vienna, Austria, 2007, pp. 267–272.
[26] A. T. Cemgil, P. H. Peeling, O. Dikmen, and S. J. Godsill, "Prior structures for time–frequency energy distributions," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, 2007, pp. 151–154.
[27] A. T. Cemgil and O. Dikmen, "Conjugate gamma Markov random fields for modelling nonstationary sources," in *Independent Component Analysis and Signal Separation*. Berlin, Heidelberg, Germany: Springer-Verlag, 2007, pp. 697–705.
[28] O. Dikmen and A. T. Cemgil, "Inference and parameter estimation in gamma chains," Dept. Eng., Univ. Cambridge, Cambridge, U.K., 2008, Tech. Rep. CUED/F-INFENG/TR.596.
[29] L. Benaroya, R. Gribonval, and F. Bimbot, "Non negative sparse representation for Wiener based source separation with a single sensor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2003, vol. 6, pp. 613–616.
[30] C. Févotte, N. Bertin, and J.-L. Durrieu, ""Nonnegative matrix factorization with the Itakura–Saito divergence with application to music analysis," *Neural Comput.*, vol. 21, pp. 793–830, Mar. 2009.

Paul H. Peeling received the M.Eng. degree from Cambridge University, Cambridge, U.K., in 2006.

In 2006, he spent three months at ARM working on the statistical modeling of hardware components. Since 2006, he has been a Research Student with the Signal Processing and Communications Laboratory, Cambridge University. His research interests include Bayesian music signal modeling and inference.

A. Taylan Cemgil (M'04) received the B.Sc. and M.Sc. degrees in computer engineering from Boğaziçi University, Istanbul, Turkey, and the Ph.D. degree from Radboud University, Nijmegen, The Netherlands, with a thesis on Bayesian music transcription.

He worked as a Postdoctoral Researcher at the University of Amsterdam and as a Research Associate at the Signal Processing and Communications Laboratory, University of Cambridge, Cambridge, U.K. He is currently an Assistant Professor at Boğaziçi University, where he cultivates his interests in machine learning methods, stochastic processes, and statistical signal processing. His research is focused towards developing computational techniques for audio, music, and multimedia processing.

Simon J. Godsill (M'95) is Professor of Statistical Signal Processing in the Engineering Department, Cambridge University, Cambridge, U.K. He has research interests in Bayesian and statistical methods for signal processing, Monte Carlo algorithms for Bayesian problems, modeling and enhancement of audio and musical signals, tracking, and high-frequency financial data. He has published extensively in journals, books, and conferences.

Prof. Godsill has acted as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the journal *Bayesian Analysis*, and as a member of IEEE Signal Processing Theory and Methods Committee. He has coedited in 2002 a special issue of the IEEE TRANSACTIONS ON SIGNAL PROCESSING on Monte Carlo Methods in Signal Processing and organized many conference sessions on related themes. He is currently co-organizing a year-long program on Sequential Monte Carlo Methods at the SAMSI Institute in North Carolina.