Emmanouil Benetos and Simon Dixon
Centre for Digital Music,
School of Electronic Engineering and
Computer Science
Queen Mary University of London
Mile End Road, London E1 4NS, UK
{emmanouilb, simond}@eecs.qmul.ac.uk

A Shift-Invariant Latent Variable Model for Automatic Music Transcription

Abstract: In this work, a probabilistic model for multiple-instrument automatic music transcription is proposed. The model extends the shift-invariant probabilistic latent component analysis method, which is used for spectrogram factorization. Proposed extensions support the use of multiple spectral templates per pitch and per instrument source, as well as a time-varying pitch contribution for each source. Thus, this method can effectively be used for multiple-instrument automatic transcription. In addition, the shift-invariant aspect of the method can be exploited for detecting tuning changes and frequency modulations, as well as for visualizing pitch content. For note tracking and smoothing, pitch-wise hidden Markov models are used. For training, pitch templates from eight orchestral instruments were extracted, covering their complete note range. The transcription system was tested on multiple-instrument polyphonic recordings from the RWC database, a Disklavier data set, and the MIREX 2007 multi-F0 data set. Results demonstrate that the proposed method outperforms leading approaches from the transcription literature, using several error metrics.

Automatic music transcription refers to the process of converting musical audio, usually a recording, into some form of notation, e.g., sheet music, a MIDI file, or a "piano-roll" representation. It has applications in music information retrieval, computational musicology, and the creation of interactive music systems (e.g., real-time accompaniment, automatic instrument tutoring). The transcription problem can be separated into several subtasks, including multi-pitch estimation (which is considered to be the core problem of transcription), onset/offset detection, instrument identification, and rhythmic parsing. Although the problem of transcribing a monophonic recording is considered to be a solved problem in the literature, the creation of a transcription system able to handle polyphonic music produced by multiple instruments remains open. For reviews on multi-pitch detection and automatic transcription approaches, the reader is referred to de Cheveigné (2006) and Klapuri and Davy (2006).

Approaches to transcription have used probabilistic methods (e.g., Kameoka, Nishimoto, and Sagayama 2007; Emiya, Badeau, and David 2010), audio feature-based techniques (e.g., Ryynänen and Klapuri 2008; Saito et al. 2008; Cañadas-Quesada et al. 2010), or machine learning approaches (e.g., Poliner and Ellis 2007). More recently, transcription systems using spectrogram-factorization techniques have been proposed (e.g., Mysore and Smaragdis 2009; Dessein, Cont, and Lemaitre 2010; Grindlay and Ellis 2010; Fuentes, Badeau, and Richard 2011). The aim of these techniques is to decompose the input spectrogram into matrices denoting spectral templates and pitch activations. Transcription systems or pitch-tracking methods that use spectrogram-factorization models similar to the ones used in this article are detailed in the following section.

Transcription approaches that use the same data sets used in this work include Poliner and Ellis (2007), where a piano-only transcription algorithm is proposed using support vector machines for note classification. For note smoothing, those authors fed the output of the classifier as input to a hidden Markov model (HMM) (Rabiner 1989). They performed experiments on a set of ten Disklavier recordings, which are also used in this article. The same postprocessing method was also used in the work of Cañadas-Quesada et al. (2010), where the joint multi-pitch estimation algorithm consists of a weighted Gaussian spectral distance measure.

Saito et al. (2008) proposed an audio feature-based multiple-F0 estimation method that uses the inverse Fourier transform of the linear power spectrum with log-scale frequency, which is called *specmurt*. The input log-frequency spectrum is considered to be generated by a convolution of a single pitch template with a pitch indicator function. The deconvolution

Computer Music Journal, 36:4, pp. 81–94, Winter 2012

of the spectrum by the pitch template results in the estimated pitch indicator function. This method is roughly equivalent to the single-component shift-invariant probabilistic latent component analysis method (Smaragdis, Raj, and Shashanka 2008), which will be detailed in the following section. Finally, we proposed an audio feature-based method for transcription (Benetos and Dixon 2011a), where joint multi-pitch estimation is performed using a weighted score function primarily based on features extracted from the harmonic envelopes of pitch candidates. Postprocessing is applied using conditional random fields.

In this article, we propose a system for polyphonic music transcription based on a convolutive probabilistic model, which extends the shift-invariant probabilistic latent component analysis model (Smaragdis, Raj, and Shashanka 2008). The original model was proposed for relative pitch-tracking (estimating pitch changes on a relative scale) using a single pitch template per source. Here, the model is proposed for multi-pitch detection, supporting the use of multiple templates per pitch and instrument source. In addition, the source contribution is time-varying, making the model more robust for transcription, and sparsity is also enforced in order to further constrain the solution. Note smoothing is performed using HMMs trained on MIDI data from the Real World Computing (RWC) database (Goto et al. 2003). The output of the system is a pitch activity matrix in MIDI units and a time-pitch representation; the latter can be used for visualizing pitch content. We presented preliminary results using the proposed model in Benetos and Dixon (2011c), where the use of a residual template was not supported and the HMM postprocessing step did not include a smoothing parameter. This article contains experiments using additional recordings from the RWC database beyond the set we used in Benetos and Dixon (2011c). Here, we present results using 17 excerpts from the RWC database (classic and jazz recordings) (Goto et al. 2003), 10 recordings from a Disklavier piano (Poliner and Ellis 2007), and the MIREX 2007 multi-F0 woodwind recording (MIREX 2007). We have performed evaluations using several error metrics from the transcription literature, and results show that the proposed model outperforms other transcription methods from the literature. This model, using a time-frequency representation with lower frequency resolution, was publicly evaluated in MIREX 2011, where the submitted system ranked second in the note-tracking task (Benetos and Dixon 2011b). Finally, the proposed model can be further expanded for musical instrument identification in polyphonic music and can also be useful in instrument-specific transcription applications.

The remainder of the article presents the shift-invariant probabilistic latent component analysis method, the proposed model, and evaluation results compared with other state-of-the-art transcription methods.

Related Work

In this section, work on automatic music transcription, pitch-tracking, and music signal analysis using probabilistic latent component analysis-based techniques will be presented in detail.

PLCA

Probabilistic latent component analysis (PLCA) is a spectrogram-factorization technique that was proposed by Smaragdis, Raj, and Shashanka (2006). It provides a probabilistic framework that is extensible as well as easy to interpret. It approximates the input spectrogram as a probability distribution $P(\omega, t)$, where ω is the frequency index and t the time index, and attempts to factorize $P(\omega, t)$ as a series of spectral components and the time activations of the respective components.

There are two forms of PLCA: asymmetric and symmetric (Shashanka, Raj, and Smaragdis 2008). Smaragdis et al. (2006) formulate the asymmetric PLCA model as:

$$P(\omega, t) = P(t) \sum_{z} P(\omega|z) P(z|t) \tag{1}$$

where $P(\omega|z)$ are the spectral templates corresponding to component z, $P(z|t)$ are the time-varying component activations, and $P(t)$ is the energy distribution of the spectrogram, which is known from the input data. For estimating $P(\omega|z)$ and $P(z|t)$, iterative update rules are used, which are derived from the Expectation-Maximization algorithm (Dempster, Laird, and Rubin 1977). It should be noted that the symmetric PLCA model decomposes $P(\omega, t)$ into $P(\omega|z)$, $P(z)$, and $P(t|z)$ (instead of the $P(z|t)$ in the asymmetric model). The symmetric model, however, is less useful when trying to control the number of components in a time frame.

Grindlay and Ellis (2010) extended the asymmetric PLCA model for polyphonic music transcription, supporting multiple spectral templates for each pitch and multiple instruments. They introduced the concept of *eigeninstruments*, which models instrument templates as mixtures of basic models in a training step. Sparsity was enforced on the transcription matrix and the source contribution matrix of the model by a tempering-based approach. For experiments, stored pitch templates from various synthesized instrument sounds were used. Experiments were performed on instrument pairs taken from the multi-track woodwind recording used in the MIREX multi-F0 development set (MIREX 2007), as well as on three J. S. Bach duets.

Mysore (2010) incorporated temporal constraints into the PLCA framework, using HMMs (Rabiner 1989). The algorithm, called the *non-negative hidden Markov model*, attempts to model the pitch changes in a monophonic recording. Each hidden state corresponds to a pitch, and multiple pitch templates are supported. Parameter estimation can be achieved using the PLCA update rules combined with the HMM forward-backward procedure. An extension for two sources was also proposed, which employed factorial HMMs.

Shift-Invariant PLCA

Smaragdis, Raj, and Shashanka (2008) extended the PLCA model to extract shifted structures in non-negative data. The algorithm, called shift-invariant PLCA, is useful for music signal processing when used with a log-frequency representation as an input, because the inter-harmonic spacings are the same for all periodic sounds. Thus, it can be used for pitch extraction and tracking. The shift-invariant PLCA model is defined as:

$$P(\omega, t) = \sum_z P(z)P(\omega|z) *_\omega P(f, t|z) \tag{2}$$

where f is the pitch-shifting factor and z is the component index. The spectral template $P(\omega|z)$ is shifted across ω, producing the time-varying pitch impulse distribution $P(f, t|z)$. $P(z)$ denotes the component prior. The EM algorithm can again be utilized for estimating the unknown parameters. By removing the convolution operator, the shift-invariant PLCA model can be expressed as:

$$P(\omega, t) = \sum_z P(z) \sum_f P(\omega - f|z)P(f, t|z) \tag{3}$$

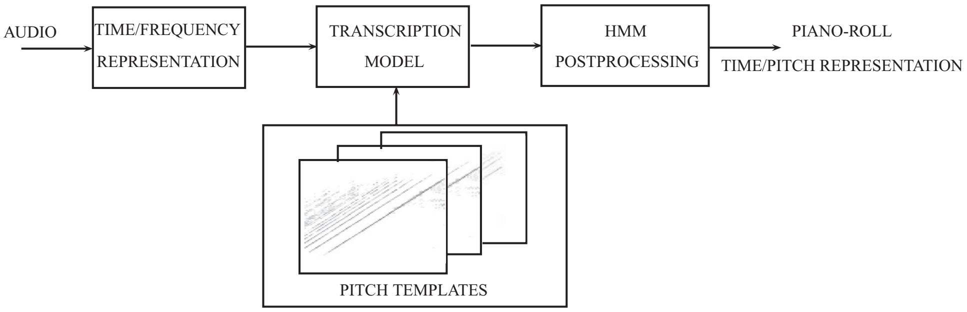
Smaragdis (2009) used the model of Equation 2 for relative pitch-tracking, where sparsity using an entropic prior was also incorporated into the model.

The shift-invariant PLCA model was utilized for multiple-instrument relative pitch-tracking by Mysore and Smaragdis (2009), with additional constraints. Firstly, a sliding Gaussian Dirichlet prior distribution was used in the computation of $P(f, t|z)$ in order to eliminate any octave errors. In addition, a Kalman filter-type smoothing is applied to $P(f, t|z)$ in order to favor temporal continuity. The method was tested on the MIREX woodwind quintet using mixtures of two instruments at a time. Fuentes, Badeau, and Richard (2011) extended the shift-invariant PLCA algorithm to detect harmonic spectra in single-pitch estimation experiments. A note was decomposed as a weighted sum of narrow-band basic harmonic spectra, and an asymmetric minimum variance prior was also incorporated into the parameter update rules in order to further constrain the model.

Proposed Method

Our goal is to propose a transcription model which expands PLCA techniques and is able to support the use of multiple spectral templates per pitch, as well as per musical instrument. In addition, the model should also be able to exploit

Figure 1. Diagram for the proposed polyphonic transcription system.


AUDIO
TIME/FREQUENCY REPRESENTATION
TRANSCRIPTION MODEL
HMM POSTPROCESSING
PIANO-ROLL TIME/PITCH REPRESENTATION
PITCH TEMPLATES

shift-invariance across log-frequency for detecting tuning changes and frequency modulations, unlike other PLCA- and non-negative matrix factorization-based transcription approaches (Grindlay and Ellis 2010; Dessein et al. 2010). Finally, the contribution of each source should be time- and pitch-dependent, contrary to the relative pitch-tracking method of Mysore and Smaragdis (2009). A diagram of the proposed transcription system can be seen in Figure 1.

Formulation

The model takes as input a log-frequency spectrogram $V_{\omega,t}$ and approximates it as a joint time-frequency distribution $P(\omega, t)$. This distribution can be expressed as a factorization of the spectrogram energy $P(t)$ (which is known) and the conditional distribution over the log-frequency bins $P_t(\omega) = P(\omega|t)$. By introducing p as a latent variable for pitch, the model can be expressed as:

$$V_{\omega,t} \approx P(\omega, t) = P(t) \sum_{p} P_t(\omega|p) P_t(p) \tag{4}$$

which is similar to the standard PLCA model, albeit with time-dependent observed spectra. By additionally introducing latent variables for instrument sources and for pitch shifting across log-frequency, the proposed model can be formulated as:

$$V_{\omega,t} \approx P(\omega, t) = P(t) \sum_{p,s} P(\omega|s, p) *_{\omega} P_t(f|p) P_t(s|p) P_t(p) \tag{5}$$

where p is the pitch index, s denotes the instrument source, and f the shifting factor. In Equation 5, $P(\omega|s, p)$ denotes the spectral templates for a given pitch and instrument source, and $P_t(f|p)$ is the time-dependent log-frequency shift for each pitch, convolved with $P(\omega|s, p)$ across ω. $P_t(s|p)$ is the time-dependent source contribution for each pitch, and $P_t(p)$ is the time-dependent pitch contribution, which can be viewed as the transcription matrix.

By removing the convolution operator in Equation 5, the model becomes:

$$P(\omega, t) = P(t) \sum_{p,f,s} P(\omega - f|s, p) P_t(f|p) P_t(s|p) P_t(p) \tag{6}$$

It should be noted that as a time-frequency representation, we use the constant-Q transform (CQT) with a spectral resolution of 120 bins/octave (Schörkhuber and Klapuri 2010). In order to utilize each spectral template $P(\omega|s, p)$ for detecting a single pitch, we constrain f to a range of one semitone. Thus, f has a length of 10.

Parameter Estimation

In order to estimate the unknown parameters in the model we use the Expectation-Maximization algorithm (Dempster, Laird, and Rubin 1977). Given the input spectrogram $V_{\omega,t}$, the log-likelihood of the model is given by:

$$\mathcal{L} = \sum_{\omega,t} V_{\omega,t} \log\big(P(\omega, t)\big) \tag{7}$$

For the *Expectation* step, we compute the contribution of latent variables p, f, s over the complete model reconstruction using Bayes' theorem:

$$P_t(p, f, s|\omega) = \frac{P(\omega - f|s, p)P_t(f|p)P_t(s|p)P_t(p)}{\sum_{p,f,s} P(\omega - f|s, p)P_t(f|p)P_t(s|p)P_t(p)} \tag{8}$$

For the *Maximization* step, we utilize the posterior of Equation 8 for maximizing the log-likelihood of Equation 7, resulting in the following update equations:

$$P(\omega|s, p) = \frac{\sum_{f,t} P_t(p, f, s|\omega + f)V_{\omega+f,t}}{\sum_{\omega,t,f} P_t(p, f, s|\omega + f)V_{\omega+f,t}} \tag{9}$$

$$P_t(f|p) = \frac{\sum_{\omega,s} P_t(p, f, s|\omega)V_{\omega,t}}{\sum_{f,\omega,s} P_t(p, f, s|\omega)V_{\omega,t}} \tag{10}$$

$$P_t(s|p) = \frac{\sum_{\omega,f} P_t(p, f, s|\omega)V_{\omega,t}}{\sum_{s,\omega,f} P_t(p, f, s|\omega)V_{\omega,t}} \tag{11}$$

$$P_t(p) = \frac{\sum_{\omega,f,s} P_t(p, f, s|\omega)V_{\omega,t}}{\sum_{p,\omega,f,s} P_t(p, f, s|\omega)V_{\omega,t}} \tag{12}$$

Equations 8–12 are iterated until convergence. By keeping the spectral templates $P(\omega|s, p)$ fixed (using pre-extracted templates in a training step), the model converges quickly, requiring about 10–20 iterations. For the present experiments, we have set the number of iterations to 15. In this work, we set $p = 1, \ldots, 89$, where the first 88 indices correspond to notes A0-C8, and the 89th index corresponds to a residual template. The spectral template update rule of Equation 9 is applied only to the 89th template, while all the other pitch templates remain fixed, unlike in Benetos and Dixon (2011c), which does not include a template update rule. The residual template is updated in order to learn the possible noise level of the recording, or any other artifacts that might occur in the music signal.

The output of the transcription model is a MIDI-scale pitch activity matrix and a pitch-shifting tensor, respectively given by:

$$\begin{aligned} P(p, t) &= P(t)P_t(p) \\ P(f, p, t) &= P(t)P_t(p)P_t(f|p) \end{aligned} \tag{13}$$

By stacking together slices of $P(f, p, t)$ for all pitch values, we can create a time–pitch representation that has a pitch resolution of 10 cents:

$$P(f', t) = [P(f, 21, t) \cdots P(f, 108, t)] \tag{14}$$

where $f' = 1, \ldots, 880$. The time–pitch representation $P(f', t)$ is useful for pitch content visualization and for the extraction of tuning information.

In Figure 2, the pitch activity matrix $P(p, t)$ for an excerpt of a guitar recording from the RWC database can be seen, along with the corresponding pitch ground truth. Also, in Figure 3, the time-pitch representation $P(f', t)$ of an excerpt of the RWC MDB-C-2001 No. 12 (string quartet) recording is shown, where vibrati in certain notes are visible. It should be noted that these vibrati would not be captured in a non-shift-invariant model.

Sparsity Constraints

Because the proposed model in its unconstrained form is overcomplete (i.e., it contains more information than in the input), especially due to the presence of the convolution operator, it would be useful to enforce further constraints in order to regulate the potential increase of information from input to output (Smaragdis 2009). To that end, sparsity is enforced on the piano-roll matrix $P_t(p)$ and the source contribution matrix $P_t(s|p)$. This can be explained intuitively, because we expect that for a given time frame only few notes should be active, whereas each pitch for a time frame is produced from typically few instrument sources.

Smaragdis (2009) enforced sparsity in the shift-invariant PLCA model by using an entropic prior, whereas Grindlay and Ellis (2010) applied a scaling factor to select update equations, which was also shown to be useful. Here, we resort to the technique of Grindlay and Ellis, which is intuitive, simpler, and easier to control. Essentially, Equations 11

Figure 2. (a) The pitch activity matrix $P(p, t)$ for the first 23 sec of RWC MDB-J-2001 No. 9 (guitar). (b) The pitch ground truth for the same recording.

Figure 3. The time–pitch representation $P(f', t)$ of the first 23s of RWC MDB-C-2001 No. 12 (string quartet). The vibrato produced in certain notes (e.g., around the 10-sec marker) can be seen.

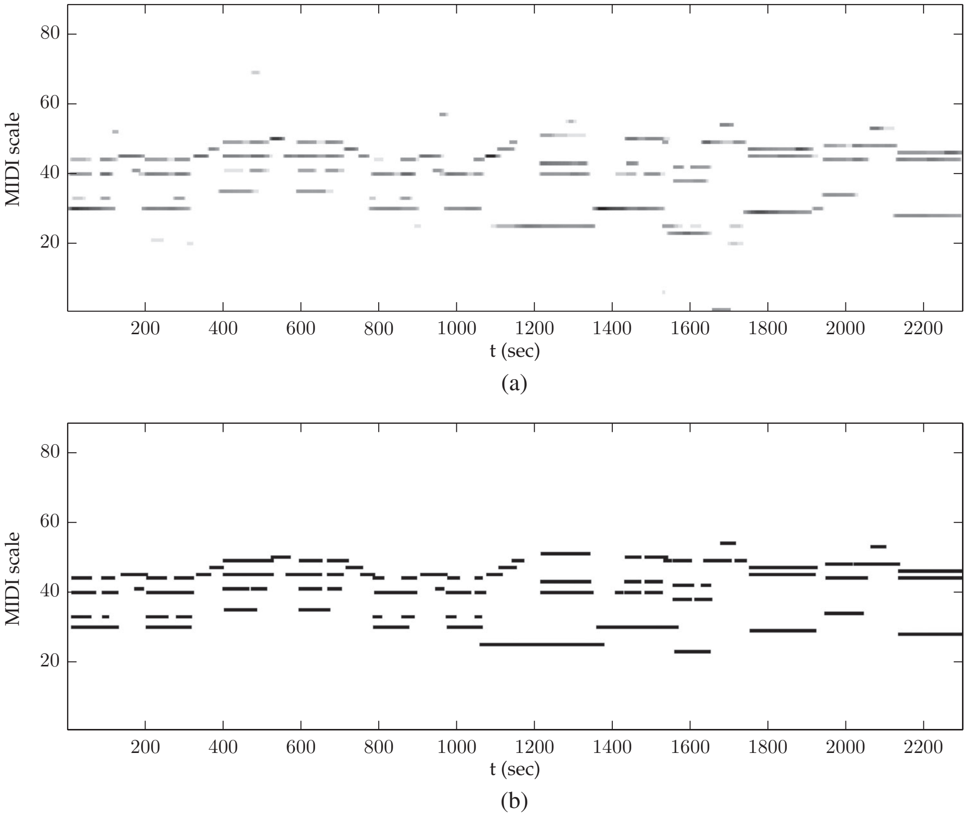


Figure 2

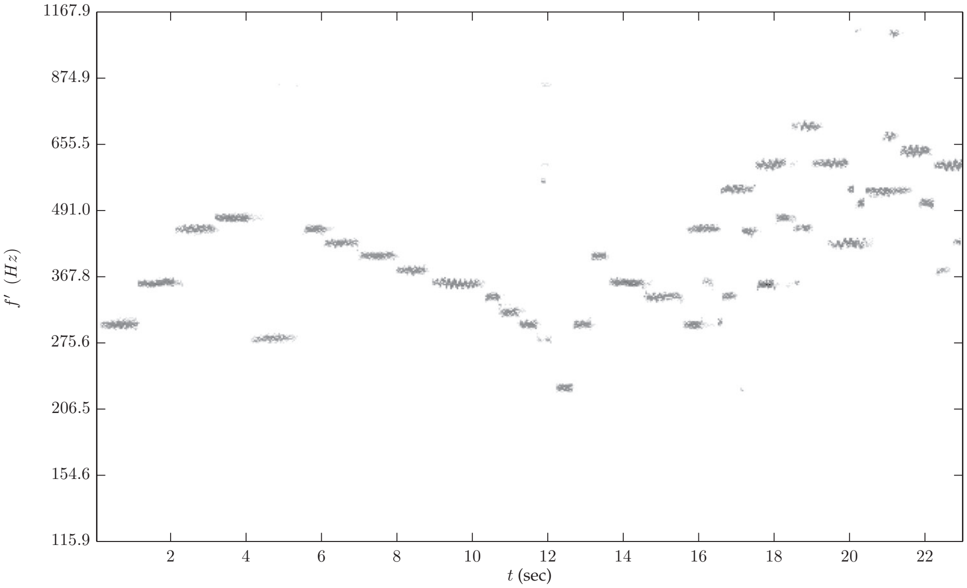


Figure 3

and 12 are modified as follows:

$$P_t(s|p) = \frac{\left(\sum_{\omega,f} P_t(p,f,s|\omega)V_{\omega,t}\right)^{\alpha}}{\sum_s \left(\sum_{\omega,f} P_t(p,f,s|\omega)V_{\omega,t}\right)^{\alpha}} \tag{15}$$

$$P_t(p) = \frac{\left(\sum_{\omega,f,s} P_t(p,f,s|\omega)V_{\omega,t}\right)^{\beta}}{\sum_p \left(\sum_{\omega,f,s} P_t(p,f,s|\omega)V_{\omega,t}\right)^{\beta}} \tag{16}$$

As Grindlay and Ellis (2010) mention, when α and β are greater than 1, the probability distributions $P_t(s|p)$ and $P_t(p)$ are "sharpened" and their entropy is lowered. This leads to fewer weights being close to 1 and most being kept near 0, thus achieving sparsity.

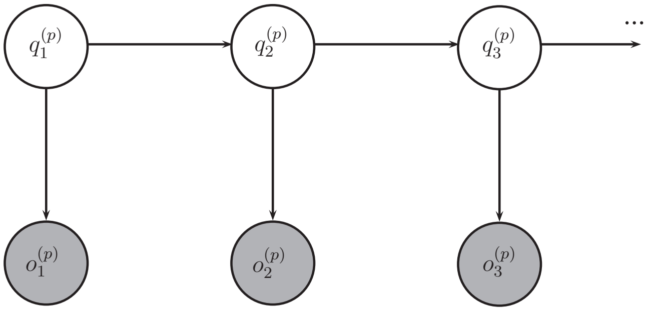
Concerning sparsity parameters, after experimentation, the sparsity for the instrument contribution matrix was set to $\alpha = 1.1$, and the sparsity coefficient for the piano-roll transcription matrix was set to $\beta = 1.3$. Although the optimal value of α when $\beta = 1$ is 1, the combination of these two parameters after experimentation yielded the optimal value of $\alpha = 1.1$.

Postprocessing

The output of spectrogram-factorization techniques for automatic transcription is typically a non-binary pitch activation matrix (e.g., see Figure 2a) which needs to be converted into a series of note events, listing onsets and offsets. Whereas most approaches extract the final note events by simply thresholding the pitch activation matrix (Dessein, Cont, and Lemaitre 2010; Grindlay and Ellis 2010), we use HMMs (Rabiner 1989) for performing note smoothing and tracking. HMMs have been used in the past for note smoothing in audio feature-based transcription approaches (e.g., Poliner and Ellis 2007; Benetos and Dixon 2011a). Here, we apply note smoothing on the pitch activity matrix $P(p,t)$.

The activity or inactivity of each pitch p is modeled by a two-state, on/off HMM. The hidden state sequence for each pitch is denoted by $Q^{(p)} = \{q_t^{(p)}\}$. MIDI files from the RWC database (Goto et al. 2003) from the classic and jazz subgenres were used in order to estimate the pitch-wise state priors $P(q_1^{(p)})$ and state transition matrices $P(q_t^{(p)}|q_{t-1}^{(p)})$. For each pitch, the most likely state sequence is given by:

Figure 4. Graphical structure of the decoding process using a pitch-wise HMM.



$$\hat{Q}^{(p)} = \arg\max_{q^{(p)}} \prod_t P\left(q_t^{(p)}|q_{t-1}^{(p)}\right) P\left(o_t^{(p)}|q_t^{(p)}\right) \tag{17}$$

which can be computed using the Viterbi algorithm (Rabiner 1989). For estimating the time-varying observation probability for each active pitch $P(o_t^{(p)}|q_t^{(p)} = 1)$, we use a sigmoid curve that has as input the piano-roll transcription matrix $P(p,t)$:

$$P\left(o_t^{(p)}|q_t^{(p)} = 1\right) = \frac{1}{1 + e^{-P(p,t)-\lambda}} \tag{18}$$

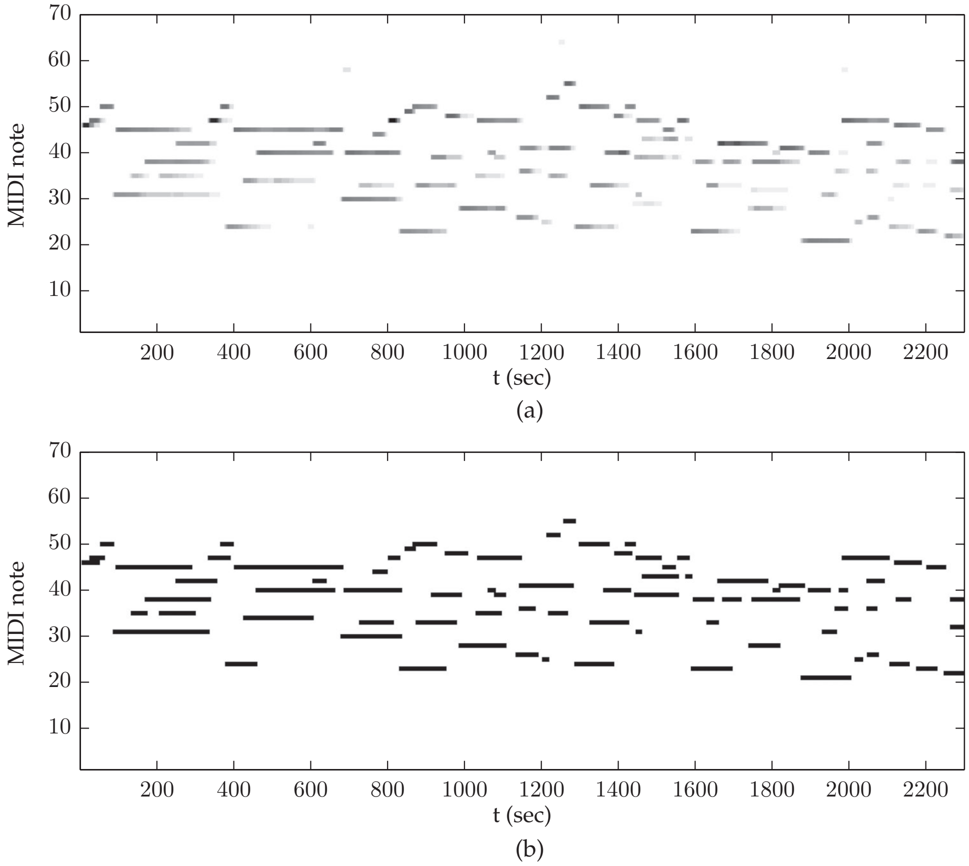
where λ is a parameter that controls the smoothing (a high value will discard pitch candidates with low probability). The graphical structure of the pitch-wise HMM decoding process can be seen in Figure 4. The result of the HMM postprocessing step is a binary piano-roll transcription, which can be used for evaluation. An example of the postprocessing step is given in Figure 5, where the transcription matrix $P(p,t)$ of a piano recording is seen along with the output of the HMM smoothing.

Training and Evaluation

Extracting Pitch Templates

Spectral templates are extracted for various orchestral instruments, using their complete note range. Isolated note samples from three different piano types were extracted from the MAPS data set

Figure 5. (a) The pitch activity matrix $P(p,t)$ of the first 23 seconds of RWC MDB-C-2001 No. 30 (piano). (b) The piano-roll transcription matrix derived from the HMM postprocessing step.

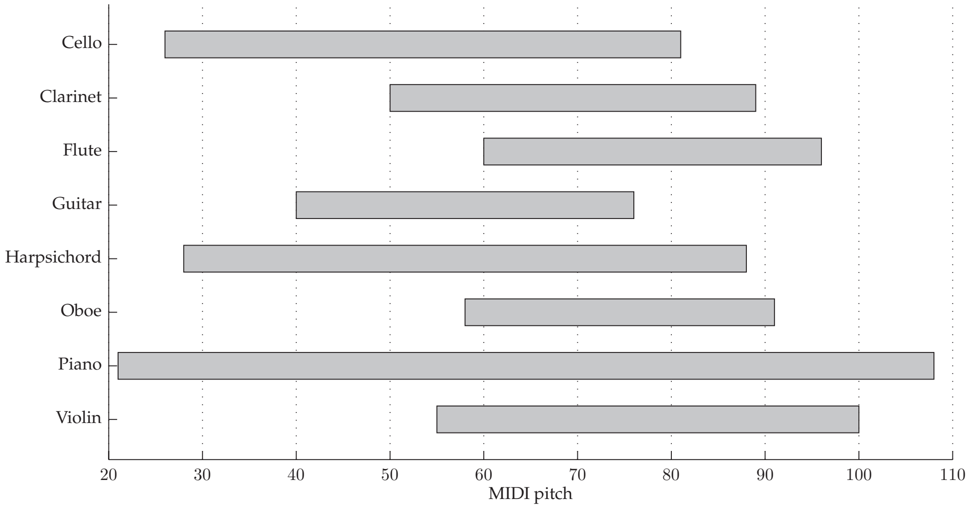


(Emiya, Badeau, and David 2010) and templates from other orchestral instruments were extracted from recordings of chromatic scales from the RWC Musical Instrument Samples data set (Goto et al. 2003), resulting in ten sets of templates, $s = 1, \ldots, 10$. The standard PLCA model of Equation 1 using only one component z was used in order to extract a single spectral template. In Figure 6, the pitch range of each instrument used for template extraction is shown.

Data Sets

For the transcription experiments, we used the set of twelve classic and jazz music excerpts from the RWC database. This data set has been used in previous research (Kameoka, Nishimoto, and Sagayama 2007; Saito et al. 2008; Cañadas-Quesada et al. 2010; Benetos and Dixon 2011a). The instruments present in these recordings are piano, guitar, flute, and bowed strings. For the track numbers, the reader can refer to Cañadas-Quesada et al. (2010). We used an additional set of five pieces from the RWC database, using the syncRWC annotations, which was evaluated in Benetos and Dixon (2011a) and that contain recordings from strings, harpsichord, and clarinet (denoted as RWC recordings 13–17). The full wind quintet recording from the MIREX multi-F0 development set (MIREX 2007) was also used for experiments. Finally, the test data set developed by Poliner and Ellis (2007) was also used for transcription experiments. It contains ten one-minute classical recordings from a Yamaha Disklavier grand piano, sampled at 8 kHz along with aligned MIDI ground truth.

Figure 6. MIDI note ranges of the instrument templates used in the proposed transcription system.



Evaluation Metrics

For the recordings used for the transcription experiments, several frame-based and note-based metrics are employed. It should be noted that frame-based evaluations take place by comparing the transcribed output and the ground-truth MIDI files at a 10-msec scale, as is the standard for the multiple-F0 MIREX evaluation (MIREX 2007). As in Grindlay and Ellis (2010), Dessein et al. (2010), and Benetos and Dixon (2011c), results are presented selecting the parameter value (in this case λ) that maximizes the average accuracy in a data set. As in Grindlay and Ellis (2011), the system is quite robust for different values of the postprocessing parameter, which can also be seen in the public evaluation results of the proposed method in MIREX 2011, using an unknown data set. For the specific experiments, the value of λ that maximizes the average accuracy is 1.2. The first frame-based metric that is used is the overall accuracy, defined by Dixon (2000):

$$Acc_1 = \frac{tp}{fp + fn + tp} \tag{19}$$

Where tp, fp, and fn refer to the number of true positives, false positives, and false negatives respectively, for all frames of the recording.

A second frame-based accuracy measure from Kameoka, Nishimoto, and Sagayama (2007) is also used, which also includes pitch substitution errors. Let $N_{ref}[t]$ stand for the number of ground-truth pitches at frame t, $N_{sys}[t]$ the number of detected pitches, and $N_{corr}[t]$ the number of correctly detected pitches. The number of false negatives at the current frame is $N_{fn}[t]$, the number of false positives is $N_{fp}[t]$, and the number of substitution errors is given by $N_{subs}[t] = \min(N_{fn}[t], N_{fp}[t])$. The accuracy measure is defined as:

$$\begin{aligned} Acc_2 &= \frac{\sum_t N_{ref}[t] - N_{fn}[t] - N_{fp}[t] + N_{subs}[t]}{\sum_t N_{ref}[t]} \\ &= \frac{\sum_t N_{ref}[t] - \max(N_{fn}[t], N_{fp}[t])}{\sum_t N_{ref}[t]} \end{aligned} \tag{20}$$

From the aforementioned definitions, several frame-based error metrics have been defined in Poliner and Ellis (2007) that measure the substitution errors (E_{subs}), missed detection errors (E_{fn}), false alarm errors (E_{fp}), and the total error (E_{tot}):

$$\begin{aligned} E_{subs} &= \frac{\sum_t \min(N_{ref}[t], N_{sys}[t]) - N_{corr}[t]}{\sum_t N_{ref}[t]} \\ E_{fn} &= \frac{\sum_t \max(0, N_{ref}[t] - N_{sys}[t])}{\sum_t N_{ref}[t]} \\ E_{fp} &= \frac{\sum_t \max(0, N_{sys}[t] - N_{ref}[t])}{\sum_t N_{ref}[t]} \\ E_{tot} &= E_{subs} + E_{fn} + E_{fp} \end{aligned} \tag{21}$$

Table 1. Transcription Results (Acc_2, in Percent) for the Twelve RWC Recordings

		Benetos and	*Cañadas-Quesada*	*Saito*	*Kameoka*
Data	*Proposed*	*Dixon (2011c)*	*et al. (2010)*	*et al. (2008)*	*et al. (2007)*
1	65.9	64.3	63.5	59.0	64.2
2	66.7	70.5	72.1	63.9	62.2
3	66.2	70.3	58.6	51.3	63.8
4	67.3	67.0	79.4	68.1	77.9
5	61.0	66.9	55.6	67.0	75.2
6	78.1	71.7	70.3	77.5	81.2
7	67.3	67.0	49.3	57.0	70.9
8	63.6	67.7	64.3	63.6	63.2
9	49.7	51.9	50.6	44.9	43.2
10	76.9	55.3	55.9	48.9	48.1
11	57.2	57.1	51.1	37.0	37.6
12	30.4	30.4	38.0	35.8	27.5
Mean	**62.5**	**61.7**	**59.1**	**56.2**	**59.6**

Table 2. Transcription Metrics (in Percent) for the Twelve RWC Recordings

Method	$\mathcal{F}$	Acc_1	Acc_2	E_{tot}	E_{subs}	E_{fn}	E_{fp}
Proposed	51.7	61.7	62.5	37.5	9.2	17.7	10.6
Benetos and Dixon (2011c)	45.2	60.8	61.7	38.3	8.9	19.6	9.8

For note-based evaluation, the system is required to return a list of notes where each note is designated by its pitch, onset time, and offset time. We utilized the onset-based metric defined in Bay, Ehmann, and Downie (2009), which is also used in the MIREX note tracking task (MIREX 2007). A note event is assumed to be correct if its onset is within ± msec of a ground-truth onset. For this case, precision, recall, and F-measure metrics are defined:

$$\mathcal{P} = \frac{N_{tp_n}}{N_{sys_n}} \quad \mathcal{R} = \frac{N_{tp_n}}{N_{ref_n}} \quad \mathcal{F} = \frac{2\mathcal{R}\mathcal{P}}{\mathcal{R}+\mathcal{P}} \tag{22}$$

where N_{tp_n} is the number of correctly detected notes, N_{ref_n} is the number of reference notes, and N_{sys_n} is the number of detected notes.

Results

Transcription results using the twelve excerpts from the RWC database and the complete set of pitch templates are shown in Table 1, compared with other state-of-the-art methods (Kameoka et al. 2007; Saito et al. 2008; Cañadas-Quesada et al. 2010; Benetos and Dixon 2011c). Additional metrics for the same experiment are presented in Table 2. The proposed model outperforms all other systems, including the PLCA-based system of Benetos and Dixon (2011c), which did not include residual basis adaptation and the smoothing parameter for the postprocessing step. Most of the errors in the present system are composed of missed detections (i.e., false negatives), whereas the number of false alarms (i.e., false positives) is significantly smaller. This means that the present system mostly detects correct pitches, but might under-determine the polyphony level. Although at first accuracy rates of about 60 percent might seem small, it should be noted that the metrics we use also take into account note durations. In fact, most of the missed detections stem from failing to detect activity in the decay part of produced notes. For note-based

Table 3. Mean Transcription Results (Acc_1, in Percent) for the Piano Recordings from Poliner and Ellis (2007)

Method	*Proposed*	*Benetos and Dixon (2011c)*	*Poliner and Ellis (2007)*	*Ryynänen and Klapuri (2005)*
Acc_1	58.9	57.6	56.5	41.2

Table 4. Transcription Metrics (in Percent) for the Piano Recordings from Poliner and Ellis (2007)

Method	$\mathcal{F}$	Acc_1	Acc_2	E_{tot}	E_{subs}	E_{fn}	E_{fp}
Proposed	60.3	58.9	58.2	41.8	9.6	17.7	14.5
Benetos and Dixon (2011c)	55.0	57.6	56.7	43.3	10.9	16.9	15.5

Table 5. Transcription Results (Acc_2, in Percent) for RWC Recordings 13–17 and the MIREX Recording

	Proposed	*Benetos and Dixon (2011c)*	*Benetos and Dixon (2011a)*
13 (RWC-MDB-C-2001 No. 13)	58.5	53.0	48.2
14 (RWC-MDB-C-2001 No. 16)	50.4	46.2	41.8
15 (RWC-MDB-C-2001 No. 24a)	64.2	58.0	66.8
16 (RWC-MDB-C-2001 No. 36)	59.6	56.6	70.7
17 (RWC-MDB-C-2001 No. 38)	70.0	61.7	75.2
MIREX multi-F0	47.1	48.1	41.3
Mean	**58.3**	**53.9**	**57.4**

metrics, the achieved F-measure for the proposed system is 51.7 percent, with $\mathcal{P} = 56.6$ percent and $\mathcal{R} = 49.3$ percent.

When viewing specific cases of recordings in Table 1 it can be seen that the best performance of the system is reported for row 6, which is a guitar recording, and row 10, which is a string quartet recording. The lowest accuracy is reported for the twelfth recording, which is a vocal performance accompanied by piano. The lower result can be explained by the fact that no pitch templates were extracted for singing voice.

Results using the Disklavier recordings from Poliner and Ellis (2007) are displayed in Table 3, compared with results from other approaches reported in Poliner and Ellis (2007) and the method in Benetos and Dixon (2011c). Because the data set consists of solo piano recordings, only the three sets of piano templates were used in the model. The proposed system again outperforms all other approaches using Acc_1. It should also be noted that the method presented in Poliner and Ellis (2007) was trained on piano data from the same source as in the test set, whereas in our case the training data were extracted from the data set in Emiya, Badeau, and David (2010). Additional metrics for the Disklavier recordings can be seen in Table 4, where it is also seen that the number of missed detections is greater than the number of false positives, although the difference this time is smaller. Regarding note-based metrics for the proposed system, $\mathcal{F} = 60.3$ percent, $\mathcal{P} = 65.5$ percent, and $\mathcal{R} = 56.5$ percent.

Finally, results using the proposed system using the five syncRWC pieces (Benetos and Dixon 2011a) and the MIREX multi-F0 woodwind quintet (MIREX 2007) can be seen in Table 5. For the

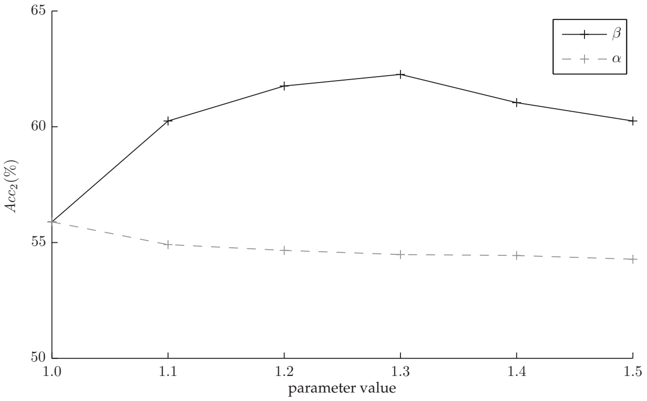


Figure 7. Transcription results (Acc_2) for RWC recordings 1–12 using various sparsity parameters (while the other parameter is set to 1.0).

five-track MIREX recording, transcription results were previously published in Mysore and Smaragdis (2009) and Grindlay and Ellis (2010), but only using pairs of these tracks. Here, results are presented for the complete mix. It should be noted that when comparing the performance of the proposed system with the one in Benetos and Dixon (2011a), the accuracy difference is 0.9 percent. The present system, however, exhibits a standard deviation of 8.5 percent compared with 15.3 percent of the system in Benetos and Dixon (2011a), demonstrating the greater robustness of the proposed model. For note-based metrics, the proposed system reaches $\mathcal{F} = 55.2$ percent for the five syncRWC pieces and $\mathcal{F} = 51.2$ percent for the MIREX recording. Regarding sparsity parameters, in Figure 7, accuracy rates for different sparsity values (α and β) are shown for RWC recordings 1–12, where the other sparsity parameter is set to 1.0.

The model of Equation 5 was publicly evaluated in the MIREX 2011 contest (Benetos and Dixon 2011b). For computational speed purposes, the CQT resolution had 60 bins/octave and fewer iterations were utilized in the update rules. Still, the submitted system ranked second in the multiple-instrument note tracking task, exhibiting high rates for the note onset metrics.

Regarding the effect of the shift-invariant model compared to a non-shift-invariant model, a comparative experiment was made in Benetos and Dixon (2011c). It was shown that the shift-invariant model outperformed the non-shift-invariant one for the twelve RWC recordings by 1.6 percent in terms of Acc_2. This difference in accuracy was mostly reported in recordings with non-ideal tuning, where semitone errors were observed in the non-shift-invariant model. Also, a comparative experiment was made using an input constant-Q transform with 60 bins/octave instead of 120. In this case, the system reaches $Acc_2 = 60.7$ percent for the twelve RWC recordings, which is a 1.8 percent decrease compared to an input CQT of 120 bins/octave. It should be noted that the proposed convolutive model can only be applied in cases where the spectral resolution is at least 2 bins/semitone.

In addition, a comparative experiment was made in order to test the effect of multiple templates for a certain instrument. The Disklavier data set of Poliner and Ellis (2007) was transcribed with the proposed system using just one set of piano templates instead of three. The resulting accuracy was $Acc_1 = 58.0$ percent, which is 0.9 percent worse compared with the set of three templates. This indicates that having several templates per instrument can help in expressing notes produced by different instrument models.

In order to test the effect of the HMM-based postprocessing step, a comparative experiment is made which replaces the smoothing procedure with simple thresholding on the pitch activity matrix $P(p,t)$. Using the set of twelve RWC recordings, the best result is $Acc_2 = 61.9$ percent, which is 0.7 percent worse compared to the HMM postprocessing step.

Concerning statistical significance, to our knowledge no statistical significance tests have been made for transcription, apart from the piecewise tests in the MIREX task (MIREX 2007) and the work done by the authors in Benetos and Dixon (2011a). In the latter, it was shown that even a small accuracy change (about 0.7 percent for the RWC data set) can be shown to be statistically significant due to the large number of data points, because transcription evaluations actually take place using 10 msec frames. Therefore the differences reported between our current work and previously published results in this section are significant.

Conclusions

This article presented a convolutive latent variable model for polyphonic music transcription, which extends the shift-invariant probabilistic latent component analysis method. The proposed model can support multiple pitch templates from multiple instruments, and can support tuning changes and frequency modulations. Unlike audio feature-based transcription systems, its architecture makes it useful for instrument-specific transcription applications, because templates from the desired instruments can easily be utilized. Moreover, the system output can be used for pitch content visualization purposes. Sparsity constraints were also enforced and note tracking was performed using HMMs. Private and public evaluation on several multiple-instrument recordings demonstrated that the proposed transcription system outperforms several state-of-the-art methods.

Future work will include an instrument identification step, which will be derived from information present in the source contribution matrix of the model and will also incorporate music signal processing-based features. Also, in order to minimize the number of missed detections observed in the present model, work will be done on addressing the amplitude modulations occurring in music signals, by modeling the temporal evolution of music sounds. Specifically, spectral templates expressing the attack, transient, sustain, and decay states of the produced notes will be used in the system, along with temporal constraints incorporated in the transcription model.

Acknowledgments

E. Benetos was funded by a Westfield Trust Research Studentship (Queen Mary University of London).

References

Bay, M., A. F. Ehmann, and J. S. Downie. 2009. “Evaluation of Multiple-F0 Estimation and Tracking Systems.” In *10th International Society for Music Information Retrieval Conference*, pp. 315–320.

Benetos, E., and S. Dixon. 2011a. “Joint Multi-Pitch Detection Using Harmonic Envelope Estimation for Polyphonic Music Transcription.” *IEEE Journal of Selected Topics in Signal Processing* 5(6):1111–1123.

Benetos, E., and S. Dixon. 2011b. “Multiple-F0 Estimation and Note Tracking Using a Convolutive Probabilistic Model.” In *Music Information Retrieval Evaluation eXchange*. Available online at www.music-ir.org/mirex/abstracts/2011/BD1.pdf. Accessed September 2012.

Benetos, E., and S. Dixon. 2011c. “Multiple-Instrument Polyphonic Music Transcription Using a Convolutive Probabilistic Model.” In *8th Sound and Music Computing Conference*, pp. 19–24.

Cañadas-Quesada, F., et al. 2010. “A Multiple-F0 Estimation Approach Based on Gaussian Spectral Modelling for Polyphonic Music Transcription.” *Journal of New Music Research* 39(1):93–107.

de Cheveigné, A. 2006. “Multiple F0 Estimation.” In D. L. Wang and G. J. Brown, eds. *Computational Auditory Scene Analysis, Algorithms and Applications*. New York: IEEE Press/Wiley, pp. 45–79.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. “Maximum Likelihood from Incomplete Data via the EM Algorithm.” *Journal of the Royal Statistical Society* 39(1):1–38.

Dessein, A., A. Cont, and G. Lemaitre. 2010. “Real-Time Polyphonic Music Transcription with Non-negative Matrix Factorization and Beta-Divergence.” In *11th International Society for Music Information Retrieval Conference*, pp. 489–494.

Dixon, S. 2000. “On the Computer Recognition of Solo Piano Music.” In *2000 Australasian Computer Music Conference*, pp. 31–37.

Emiya, V., R. Badeau, and B. David. 2010. “Multipitch Estimation of Piano Sounds using a New Probabilistic Spectral Smoothness Principle.” *IEEE Transactions on Audio, Speech, and Language Processing* 18(6):1643–1654.

Fuentes, B., R. Badeau, and G. Richard. 2011. “Adaptive Harmonic Time–Frequency Decomposition of Audio using Shift-Invariant PLCA.” In *IEEE International Conference on Audio, Speech and Signal Processing*, pp. 401–404.

Goto, M., et al. 2003. “RWC Music Database: Music Genre Database and Musical Instrument Sound Database.” In *International Conference on Music Information Retrieval*, pp. 229–230.

Grindlay, G., and D. Ellis. 2010. "A Probabilistic Subspace Model for Multi-Instrument Polyphonic Transcription." In *11th International Society for Music Information Retrieval Conference*, pp. 21–26.

Grindlay, G., and D. Ellis. 2011. "Transcribing Multi-Instrument Polyphonic Music with Hierarchical Eigeninstruments." *IEEE Journal of Selected Topics in Signal Processing* 5(6):1159–1169.

Kameoka, H., T. Nishimoto, and S. Sagayama. 2007. "A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering." *IEEE Transactions on Audio, Speech, and Language Processing* 15(3):982–994.

Klapuri, A., and M. Davy, eds. 2006. *Signal Processing Methods for Music Transcription*. New York: Springer-Verlag, 2nd edition.

MIREX. 2007. "Music Information Retrieval Evaluation eXchange (MIREX)." Available online at http://music-ir.org/mirexwiki/. Accessed September 2012.

Mysore, G. 2010. "A Non-Negative Framework for Joint Modeling of Spectral Structure and Temporal Dynamics in Sound Mixtures." PhD thesis, Stanford University, Palo Alto, CA.

Mysore, G., and P. Smaragdis. 2009. "Relative Pitch Estimation of Multiple Instruments." In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 313–316.

Poliner, G., and D. Ellis. 2007. "A Discriminative Model for Polyphonic Piano Transcription." *EURASIP Journal on Advances in Signal Processing* (8):154–162.

Rabiner, L. R. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE* 77(2):257–286.

Ryynänen, M., and A. Klapuri. 2005. "Polyphonic Music Transciption using Note Event Modeling." In *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 319–322.

Ryynänen, M., and A. Klapuri. 2008. "Automatic Transcription of Melody, Bass Line, and Chords in Polyphonic Music." *Computer Music Journal* 32(3):72–86.

Saito, S., et al. 2008. "Specmurt Analysis of Polyphonic Music Signals." *IEEE Transactions on Audio, Speech, and Language Processing* 16(3):639–650.

Schörkhuber, C., and A. Klapuri. 2010. "Constant-Q Transform Toolbox for Music Processing." In *7th Sound and Music Computing Conference*, pp. 322–329.

Shashanka, M., B. Raj, and P. Smaragdis. 2008. "Probabilistic Latent Variable Models as Nonnegative Factorizations." *Computational Intelligence and Neuroscience*. Available online at www.hindawi.com/journals/cin/2008/947438. Accessed September 2012.

Smaragdis, P. 2009. "Relative-Pitch Tracking of Multiple Arbitary Sounds." *Journal of the Acoustical Society of America* 125(5):3406–3413.

Smaragdis, P., B. Raj, and M. Shashanka. 2006. "A Probabilistic Latent Variable Model for Acoustic Modeling." In *Neural Information Processing Systems Workshop* (pages unnumbered).

Smaragdis, P., B. Raj, and M. Shashanka. 2008. "Sparse and Shift-Invariant Feature Extraction from Non-Negative Data." In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2069–2072.