J Intell Inf Syst (2013) 40:141–158
DOI 10.1007/s10844-012-0220-9

Instrument identification and pitch estimation in multi-timbre polyphonic musical signals based on probabilistic mixture model decomposition

Ying Hu · Guizhong Liu

Received: 9 April 2012 / Revised: 12 August 2012 / Accepted: 14 August 2012 /
Published online: 1 September 2012

Abstract In this paper, we propose a method based on probabilistic mixture model decomposition that can simultaneously identify musical instrument types, estimate pitches and assign each pitch to its source instrument in monaural polyphonic audio containing multiple sources. In the proposed system, the probability density function (PDF) of the observed mixture note is treated as a weighted sum approximation of all possible note models. These note models, covering 14 instruments and all their possible pitches, describe their dynamic frequency envelopes in terms of probability. The weight coefficients, indicating the probabilities of the existence of pitches of a certain type of instrument, are estimated using the Expectation-Maximization (EM) algorithm. The weight coefficients are used to detect the types of source instruments and the pitches. The results of experiments involving 14 instruments within a designated pitch range F3–F6 (37 pitches) demonstrate a good discrimination capability, especially in instrument identification and instrument-pitch identification. For the entire system including the note onset detection tool, using quartet polyphonic recordings, the average F-measure values of instrument-pitch identification, instrument identification and pitch estimation were 55.4, 62.5 and 86 % respectively.

Keywords Instrument identification · Instrument-pitch identification · Pitch estimation · EM algorithm · Probabilistic model

1 Introduction

Music signals are generally composed of a collection of sounds that may contain multiple instruments and/or multiple pitches. The identification of the instruments

Y. Hu · G. Liu (✉)
Xi’an Jiaotong University, Xi’an, China
e-mail: liugz@xjtu.edu.cn

and the estimation of the pitches in a multi-timbre polyphonic recording are important tasks in music information retrieval (MIR) that can be used by other digital audio applications. Instrument identification can be used to improve musical genre classification, or to create instrument-specific rules to improve the quality of sound source separation and for instrument-specific queries. Pitch estimation can be used for chord detection, instrument identification and source separation. Transcription is one of the most widespread uses of pitch estimation. Musical instrument identification and pitch estimation have received increasing attention in recent years.

Early research in this area studied either musical instrument identification (Burred et al. 2010; Hu and Liu 2011; Essid et al. 2006; Heittola et al. 2009; Jiang et al. 2009; Joder et al. 2009; Kostek 2004; Kursa et al. 2009; Loughran et al. 2008; Wieczorkowska and Kubera 2010), or pitch estimation (Bertin et al. 2009; Rao and Shandilya 2004; Dessein et al. 2010; Goto 2004; Kameoka et al. 2007; Vincent et al. 2010). In addition, there is a need to know which musical instrument played a particular pitch in the process of transcription. Research has also investigated joint pitch estimation and instrument identification (Grindlay and Ellis 2010; Kitahara et al. 2007; Wu et al. 2011). In Wu et al. (2011), the pitches following the parameters of a note were first estimated using the Expectation-Maximization (EM) algorithm, and the timbre features were then extracted from the estimated parameters for musical instrument recognition. Thus the instrument identification was significantly affected by the results of pitch estimation. Similarly, in Kitahara et al. (2007), the probabilistic representations of the existence of an instrument were affected by the probability of the existence of the pitch in a non-specific instrument. Therefore, if the system did not perform well at pitch estimation, it would not have desirable results in instrument identification. The dynamic characteristic in a note is an important factor contributing to the perception of timbre (Burred et al. 2010; Hu and Liu 2011) and was considered for instrument identification in the methods proposed by Kitahara et al. (2007) and in our early research (Hu and Liu 2011). In contrast, the Probabilistic Eigen-instrument Transcription (PET) system presented in Grindlay and Ellis (2010) did not concern itself with the dynamic characteristic, so their PET system did not perform well at instrument identification.

Most researchers focused specifically on partials including the fundamental frequency (f_o) and its overtones. The partials are helpful in instrument identification (Barbedo and Tzanetakis 2011; Burred et al. 2010; Hu and Liu 2011), multi-timbre sound separation (Li et al. 2009; Bay and Beauchamp 2006; Dziubinski et al. 2005) and pitch estimation (Bertin et al. 2009; Goto 2004; Kameoka et al. 2007; Vincent et al. 2010), because the evolution (dynamic characteristic) of partials with time and the relationship between the partials in a note could distinguish musical instruments, while the positions of partials with respect to frequency could distinguish pitch. However, partial extraction faces challenges as follows:

- The result of f_o detection has a strong impact on partial extraction based on the fundamental frequency (Barbedo and Tzanetakis 2011; Li et al. 2009).
- Even if the detected f_o is completely correct extracted incorrect partials may exist for the inherent inharmonicity of some instruments (Barbedo and Tzanetakis 2011).
- Partial extraction based on sinusoidal-modeling, while addressing f_o-dependency, still does not have exactly the same amplitude and frequency position of the original partial because of the influence of noise (Burred et al. 2010).

Thus, despite the clear role of partials in instrument identification and pitch estimation, direct partial extraction is inadvisable. This paper presents a simple and reliable strategy to identify instruments, estimate pitch and allocate the pitch to the corresponding instrument synchronously in a multi-timbre polyphonic signal. The proposed method avoids direct partial extraction, and therefore the instrument identification is not affected by the pitch estimation.

The probabilistic model method, which is a valid approach for automated class indexing, was employed by the researchers in this area (Heittola et al. 2009; Goto 2004; Kameoka et al. 2007; Grindlay and Ellis 2010; Kitahara et al. 2007; Wu et al. 2011). In this study we continue to use this method to model magnitude spectrograms of each note, while simultaneously taking into consideration all types of musical instruments and all the possible pitches of each instrument, and treating the mixture magnitude spectrogram of each input musical note as if it contains all types of musical instruments and all possible pitches with different weights. Each weight coefficient, representing the existence probability of a pitch played by a certain type of instrument, is then estimated using the EM algorithm.

This paper is organized as follows. Section 2 describes the steps of the algorithm. Section 3 presents the experiments and corresponding results. Finally, Section 4 presents the conclusions and final remarks.

2 The algorithm

Probabilistic Latent Semantic Analysis (PLSA) was first used for automatic indexing and information retrieval of text documents (Hofmann 1999), and was then developed as the Probabilistic Latent Variable Model (PLVM) for the analysis of acoustic spectra (Smaragdis et al. 2006). The PLVM is a class of probabilistic models employing latent variables. Goto modeled the observed mixture probability as a weighted mixture of all possible tone probabilistic models for pitch estimation, which coincided with the major idea behind the PLVM (Goto 2004). Shashanka et al. (2008) have shown that the PLVM decompositions are numerically identical to Non-negative Matrix Factorization (NMF), which has been shown to be a useful approach in polyphonic music transcription (Bertin et al. 2009; Vincent et al. 2010). However, Grindlay et al. confirmed that generic NMF-based transcription performs worse than probabilistic model-based transcription (Grindlay and Ellis 2010). All this research employs the PLVM to estimate pitch, but to our knowledge, there is hardly any research employing PLVM to address the identification problem of musical instruments.

The algorithm presented in this paper primarily computes the decomposition of the probabilistic mixture model (non-negative data) that maximizes the likelihood of the approximation data given the actual (observed) data. The proposed probabilistic models contain the dynamic envelope of all frequency components in a note. More computation is required and noise exists in non-partial components, but it would not contain the errors of partial frequency and partial amplitude because it immediately avoids the extraction of partials. It is important that the system can indicate latent classes (the type of musical instrument and the pitch) simultaneously and does not need to know the number of sources in advance.

Our proposed algorithm involves the following steps: computing a time-frequency representation of the signal, dividing it into notes, obtaining probabilistic note models belonging to the pitch of each instrument and decomposing the probabilistic mixture note using the EM algorithm. The following subsections explain each step in detail.

2.1 Time-frequency representation

To discriminate each partial, the time-frequency (T-F) representation must have a resolution of at least one semitone over the whole frequency range. In the following, we consider three particular frequency scales.

The basic T-F representation is a magnitude spectrogram with linear and uniformly-spaced frequencies. A Hamming window of a 4096 point short-time Fourier transform (STFT) with a hop size of 23 ms was taken for a single-channel multi-timbre polyphonic signal with a 44.1 kHz sampling rate.

Another scale is the Musical Instrument Digital Interface (MIDI) semitone scale (referred to as the “semitone scale”) which is related to its fundamental frequency by:

$$v_{i0}^{Hz} = 440 \times 2^{\frac{p_i - 69}{12}}, \tag{1}$$

where $p_i = 21, \ldots, 116$. We selected 96 semitones (8 octaves) covering the frequency range from 27.5 Hz (pitch A0), which is the lower bound of the pitch range of a piano, to 7,040 Hz embracing all the fundamental frequencies of the piano pitch and most of its partials. The semitone scale time-representations versus the frequencies on a logarithmic scale were acquired using a constant-Q transform with $Q = 16.817$ (Brown 1991). The time resolution was set to 23 ms for all sub-bands and the frequency resolution was one semitone.

The third frequency scale is the Equivalent Rectangular Bandwidth (ERB) scale given by:

$$v_f^{ERB} = 9.26 \log\left(0.00437 v^{Hz} + 1\right) \tag{2}$$

The input signal was passed through a set of $F = 250$ gammatone filters of order 4 indexed by f with frequency v_f^{ERB} linearly spaced on the ERB scale (Vincent et al. 2010). Gammatone filter is derived from psychophysical and physiological observations of the auditory periphery, and this filter bank is a standard model of cochlear filtering. The length of each filter was set so that the bandwidth of its main frequency lobe was equal to four times the difference between its frequency and those of adjacent filters. Each sub-band was then portioned into overlapped 46 ms time frames with hop size of 23 ms indexed by t and the root-mean-square magnitude $X(f, t)$ was computed within each frame.

2.2 Note segmentation

The follow-up tasks were all based on the mixture recording of the note level, so we first divided the time-frequency representation of signal, which had a minimum time unit of a frame, into notes. The entire system included the onset detector proposed in our earlier research (Hu and Liu 2011). However, to accurately measure the performance of our proposed probabilistic mixture model decomposition (PMMD) method, it was necessary to guarantee to rule out all other errors, so one of the main tests presented in Section 3 was performed assuming the note onsets were known.

Since it is also important to know how the entire system works, a study on the effects of the error of note onset on the performance of our proposed system and the experimental results of the entire system is presented in Section 3.

2.3 Probabilistic mixture model

The time variation of the envelope should be modeled in a reliable manner, since it plays an important role when characterizing timbre. Therefore, the choice was always to retain the fixed sequence ordering of the amplitude of each frequency bin. To this end, the frame-wise envelope interpolation was introduced in which the amplitude envelopes were interpolated along the axis of the time for each frequency bin. We selected the cubic interpolation method according to the conclusions of Burred et al. (2010). Let $X(f, t)$ denote the T-F amplitude spectrum of a note corresponding to the frequency f (with $f = 1, \ldots, F$), which is one of the three above-mentioned frequency scales, and the time t (with $t = 1, \ldots, T$), the whole matrix X conveys variations of the frequency properties of sound signal over time. To retain the fixed sequence ordering, for each frequency bin, the amplitude envelope should be interpolated in time with the fixed maximum frame length limit $\mathrm{T}_{\max}$. So that the durations of the musical notes represented in each model and those of the mixture notes to be analyzed are always the same. In the following, the sign $\sim$ will denotes interpolation. $\left\{\widetilde{X}\,(f, t)|\ f = 1, \ldots, F;\ t = 1, \ldots, \mathrm{T}_{\max}\right\}$, where $\widetilde{X}\,(f, t)$ represents time-frequency spectrum that it has been interpolated over time. Note that T is the actual frame length of a note before interpolation while $T_{\max}$ is the frame length of a note after interpolation.

Considering an observed T-F amplitude spectrum $X(f, t)$ was generally a mixture note of multi instruments and/or multi pitches. We assumed the observed magnitude spectrum was approximately the weighted sum of the independent magnitude spectra generated from the underlying sources (latent variables). Figure 1 is a graphical representation of the mixture of two sources, where ω_1 and ω_2 are weighting coefficients of two of the sources. To enable the application of the statistical method, each component represented a probability function. The observed T-F amplitude spectrum $\widetilde{X}\,(f, t)$ can be modeled as a joint distribution over time and frequency:

$$P_{\widetilde{X}}\,(f, t) = \frac{\widetilde{X}\,(f, t)}{\sum_{f=1}^{\mathrm{F}} \sum_{t=1}^{\mathrm{T}_{\max}} \widetilde{X}\,(f, t)}, \tag{3}$$

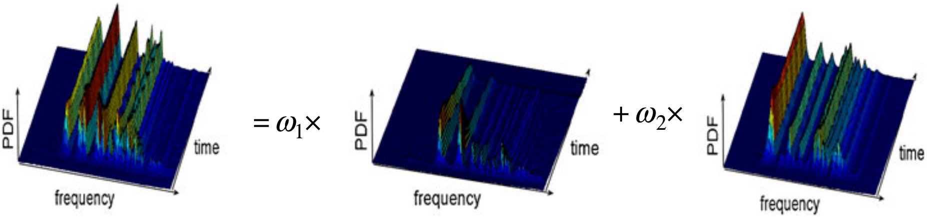

Fig. 1 The mixture of two sources

where F is the maximum frequency bin. We then considered the observed probability density function (PDF), $P_{\widetilde{X}}\,(f, t)$, which approximated to the following model $P\,(f,\ t|\,\omega)$, as a weighted mixture of all possible models $P\,(\,f, t|\ p, m)$:

$$P\,(f,\ t|\,\boldsymbol{\omega}) = \sum_{p=Pl_m}^{Ph_m} \sum_{m=1}^{\mathrm{M}} \omega\,(p, m)\ P\,(f, t\,|p, m\,), \tag{4}$$

$$\boldsymbol{\omega} = \{\omega\,(p, m)|\ 1 \le m \le M,\ Pl_m \le p \le Ph_m\}\,, \tag{5}$$

where p and m denote the pitch and instrument, M is the number of recognizable instruments, and Pl_m and Ph_m the lower and upper bounds of the allowable pitch range of instrument m. Because the numbers of instruments and pitches were unknown, it was important to simultaneously take into consideration all pitches and all instruments. In (4), there are two latent (unobserved) variables: the type of musical instrument m and the pitch p.

$P\,(\,f, t|\ p, m)$ is the probabilistic model of a note given instrument m and pitch p. It characterizes the typical spectrogram of a single source (here we refer to a certain pitch played by a certain musical instrument), that reflects the envelope of each frequency component including the entire partials along the time axis, and reveals the relative relationship between adjacent partials in the form of probability. We assumed that each pitch of a certain instrument had only one model instead of considering the diversity of the player and instrument. Specifically, the T-F representation of given instrument m and pitch p is analyzed by one of three methods mentioned in Section 2.1. After frame-wise interpolation process of amplitude envelope, we obtain the uniform T-F matrix of same size $\widetilde{X}\,(\,f, t|\ p, m)$. Then the probabilistic model $P\,(\,f, t|\ p, m)$ is got by (3). It should be said that the T-F representation method of probabilistic model is the same as that of observed mixture note. There was only one sample for each pitch of each instrument in the training dataset. Each sample had the same time-frequency representation and the same duration of the note as the mixture note to be analyzed. The probabilistic models of each note were calculated by (3) after the amplitude interpolation process. One potential issue that we had to face was the differences in the natural pitch ranges of the instruments. For example, an acoustic bass could not play above C5 while an oboe could not play below C4. Therefore, the practical models parameters $P\,(\,f, t|\ p, m)$ were performed within the pitch range of instrument m so that it would be masked out (set with 0) if the pitch p was outside the pitch range $[Pl_m, Ph_m]$.

$\omega(p, m)$ was the weight for the pitch p of the instrument m that satisfied:

$$\sum_{p=1}^{P_{\max}} \sum_{m=1}^{M} \omega\,(p, m) = 1, \quad \forall p, \forall m:\ \ 0 \le \omega\,(p, m) \le 1. \tag{6}$$

$P_{\max}$ was the number of pitches in the designated range. The unknown variables were the weight matrix $\boldsymbol{\omega}$ in (4). $\boldsymbol{\omega}$ reveals the probability of the existence of pitch p of instrument m in observed note $X(f,\ t)$. The more dominant a note model $P\,(\,f, t|\ p, m)$ was in the mixture, the higher the value of the corresponding $\boldsymbol{\omega}$. In addition, for instrument m, the weight $\boldsymbol{\omega}$ of the pitch p that was out of the pitch range $[Pl_m, Ph_m]$ was zero when the corresponding probabilistic model was also zero. Thus it could rule out illogical allocation of pitches to specific instruments.

The overall flowchart of the proposed system is illustrated in Fig. 2. After the acquisition of the probabilistic model of each pitch of each instrument, the observed probabilistic mixture should be decomposed according to (4) to acquire the parameter $\boldsymbol{\omega}$, and then to start the next step (salience measurement) for obtaining the types of musical instrument, the pitches and the pitches played by a certain musical instrument.

2.4 MAP estimation using the EM algorithm

The decomposition of the probabilistic mixture model to acquire the weights $\boldsymbol{\omega}$ is a problem of Maximum A posteriori Probability (MAP). When the magnitude spectrogram of a mixture note $X(f, t)$ was observed, after amplitude envelope interpolation one can get $\widetilde{X}(f, t)$, and then it could be further modeled as a joint distribution over time and frequency $P_{\tilde{X}}(f, t)$ by (3). The MAP estimation of $\boldsymbol{\omega}$ could be obtained by maximizing the log-likelihood function given the observed (incomplete) data:

$$\begin{aligned} L &= \iint_D P_{\tilde{X}}(f, t) \log P(f, t \,|\, \boldsymbol{\omega}) \, df dt \\ &= \iint_D P_{\tilde{X}}(f, t) \log \left(\sum_{p,m} \omega(p, m) \, P(f, t | \, p, m) \right) df dt. \end{aligned} \tag{7}$$

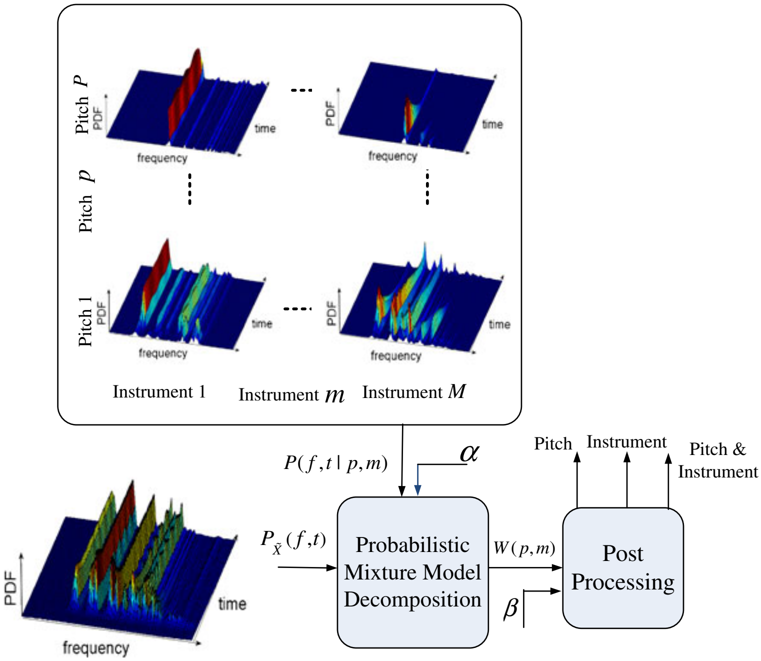

Fig. 2 Overview of the proposed probabilistic mixture model decomposition (PMMD) system

Each $\widetilde{X}(f, t)$ is defined on a domain:

$$D = \{f, t \in \mathbb{R} | 1 \leq f \leq \mathrm{F}, 1 \leq t \leq \mathrm{T}_{\max}\}. \tag{8}$$

In the case of discrete-time observations, the integral is replaced by the sum over all discrete points of f and t. However, this maximization problem was too difficult to solve analytically because it contained the log of the sum, so we subsequently employed the EM algorithm to estimate the parameter $\boldsymbol{\omega}$ (Goto 2004; Wu et al. 2011). Firstly, $\boldsymbol{\omega}$ is initialized with random values. Then, the value of the parameter set in each iteration of the EM algorithm according to the following two steps, the expectation step (E-step) and the maximization step (M-step), to compute MAP estimation from incomplete observed data until convergence. By introducing the hidden variable (unobserved data) m and p, the two steps can be specified as follows:

E-step The EM algorithm (Bilmes 1998) first finds the conditional expected value of the complete-data log-likelihood log $P(f, t, p, m | \boldsymbol{\omega})$ with respect to the hidden data p and m given the observed data f and t, and the current parameter estimates $\boldsymbol{\omega}^{(i-1)}$. The calculation of this conditional expectation is called the E-step of the algorithm. We define the weighted sum of the conditional expectation i.e. Q function:

$$Q\left(\boldsymbol{\omega}, \boldsymbol{\omega}^{(i-1)}\right) = \iint_D P_{\widetilde{X}}(f, t)\, E_{p,m}\left[\log p\,(f, t, p, m | \boldsymbol{\omega}) | f, t, \boldsymbol{\omega}^{(i-1)}\right] df dt, \tag{9}$$

where $E_{p,m}\left[a | b\right]$ denotes the conditional expectation of a with respect to hidden variables p and m, with the probability distribution determined by condition b. The PDF of observed amplitude spectrum of a note $P_{\widetilde{X}}(f, t)$ here can be considered as weighting. $\boldsymbol{\omega}^{(i-1)}$ are the current(or the i-1 th iteration)estimates of the parameters that we used to evaluate the expectation and $\boldsymbol{\omega}$ are new parameters we used to optimize the increase in Q. The conditional expectation term in the right side of (9) can be re-written as:

$$\begin{aligned} &E_{p,m}\left[\log P\,(f, t, p, m | \boldsymbol{\omega}) | f, t, \boldsymbol{\omega}^{(i-1)}\right] \\ &\quad = \sum_{p,m} P\left(p, m | f, t, \boldsymbol{\omega}^{(i-1)}\right) \log P\,(f, t, p, m | \boldsymbol{\omega}) \cdot \end{aligned} \tag{10}$$

Based on the Bayes rules, we obtain the PDF of hidden data given the observed data and current parameters:

$$P\left(p, m | f, t, \boldsymbol{\omega}^{(i-1)}\right) = \frac{\omega^{(i-1)}(p, m)\, P(f, t | p, m)}{\sum\limits_{p,m} \omega^{(i-1)}(p, m)\, P(f, t | p, m)}. \tag{11}$$

Based on the probability product rule, we obtain the PDF of the complete-data:

$$P(f, t, p, m | \boldsymbol{\omega}) = \omega(p, m)\, P(f, t | p, m). \tag{12}$$

Substituting (11) and (12) into (10), and then further substituting (10) into (9), we can obtain the exhaustive $Q(\boldsymbol{\omega}, \boldsymbol{\omega}^{(i-1)})$ function:

$$Q(\boldsymbol{\omega}, \boldsymbol{\omega}^{(i-1)}) = \iint_D P_{\widetilde{X}}(f, t) \cdot \sum_{p,m} \left[\frac{\omega^{(i-1)}(p, m)\, P(f, t\,|p, m)}{\sum\limits_{p,m} \omega^{(i-1)}(p, m)\, P(f, t|\,p, m)} (\log \omega(p, m) + \log P(f, t|\,p, m)) \right] dfdt. \tag{13}$$

M-step The second step is to update the parameters $\boldsymbol{\omega}$ by maximizing the expectation $Q\left(\boldsymbol{\omega}\,\middle|\boldsymbol{\omega}^{(i-1)}\right)$ computed in the first step.

$$\boldsymbol{\omega}^i = \underset{\boldsymbol{\omega}}{\arg\max}\; Q\left(\boldsymbol{\omega}, \boldsymbol{\omega}^{(i-1)}\right). \tag{14}$$

To find the expression for parameters of i-th iteration $\boldsymbol{\omega}^i$, we introduced the Lagrange multiplier λ with the constraint of (6), and solved the following equation:

$$\frac{\partial}{\partial \omega(p, m)} \left[Q\left(\boldsymbol{\omega}|\, \boldsymbol{\omega}^{(i-1)}\right) - \lambda \left(\sum_{p,m} \omega(p, m) - 1 \right) \right] = 0. \tag{15}$$

From (6) and (15), we found that $\lambda = 1$ resulting in:

$$\omega^i(p, m) = \iint_D P_{\widetilde{X}}(f, t) P\left(p, m|\, f, t, \boldsymbol{\omega}^{(i-1)}\right) dfdt. \tag{16}$$

To avoid unfavorable local extreme values or over-fitting, a generalization of maximum likelihood for the mixture model, named tempered EM (TEM) and based on entropic regularization, was proposed in Hofmann (1999). This introduced a control parameter α (inverse computational temperature) and modified the E-step (11) according to:

$$P\left(p, m|\, f, t, \boldsymbol{\omega}^{(i-1)}\right) = \frac{\omega^{(i-1)}(p, m)\left[P(f, t\,|p, m)\right]^{\alpha}}{\sum\limits_{p,m} \omega^{(i-1)}(p, m)\left[P(f, t\,|p, m)\right]^{\alpha}}. \tag{17}$$

We selected an α slightly larger than 1, this was similar to encouraging sparseness in probabilistic models in Grindlay and Ellis (2010). It was clear that when α was larger than 1, the $P\left(p, m|\, f, t, \boldsymbol{\omega}^{(i-1)}\right)$ distributions over p and m were "sharpened", thus encouraging sparseness that was only a few pitches, and a few instruments that were actively playing a given note.

In summary, in the process of iteration, we only needed to update the posterior distribution over the hidden variables p and m for each time-frequency point given the current estimates $\boldsymbol{\omega}^{(i-1)}$ by (17) in the E-step, and use this posterior to update the parameters $\boldsymbol{\omega}^i$ by (16) in the M-step, until convergence.

2.5 Salience measurement

After the decomposition of the probabilistic mixture model using the EM algorithm, the weights matrix $\boldsymbol{\omega}$ was acquired. It then needed a processing for salience measurement to characterize which pitches and/or which instruments were definitely active. We used a simple and efficient strategy used in Vincent et al. (2010):

$$\left\{ p|\, P\left(p\right) \geq \beta_1 \max_p \left(P\left(p\right)\right),\, P\left(p\right) = \sum_m \omega\left(p, m\right) \right\} \tag{18}$$

$$\left\{ m|\, P\left(m\right) \geq \beta_2 \max_m \left(P\left(m\right)\right),\, P\left(m\right) = \sum_p \omega\left(p, m\right) \right\} \tag{19}$$

$$\left\{ p, m|\, \omega\left(p, m\right) \geq \beta_3 \max_{p,m} \left(\omega\left(p, m\right)\right) \right\}. \tag{20}$$

Using (18) and (19), we determined the pitches and the types of instrument that exist in the observed data. From (20), not only the pitches but the types of corresponding instruments are also detected. We call this instrument-pitch identification. The salience measurements of the sources (pitch or instrument) are all based on the comparison between the probabilities of the existence of certain source and the maximum probability of its existence. The threshold $\{\beta_i | i = 1, 2, 3\}$ can either be set manually or learned from training data.

An example is shown in Fig. 3, which is a diagrammatic presentation of the weight matrix, where the observed recording mixture contains four pitches: pitch F#4 (MIDI index 66) played by an acoustic bass (‘ab’), pitch A4 (MIDI index 69) played by a violin (‘vi’), pitch D6 (MIDI index 86) played by a marimba (‘mb’) and pitch D5 (MIDI index 74) played by a celesta (‘cs’). The musical instrument index and the corresponding instrument name are shown in Table 1. A rectangular block in Fig. 3 indicates a point in weight matrix $\boldsymbol{\omega}$. The darker the rectangular block is, the higher the probability of the existence of a corresponding instrument and pitch. The salience measurement can be observed intuitively from the color depth of these rectangular blocks.

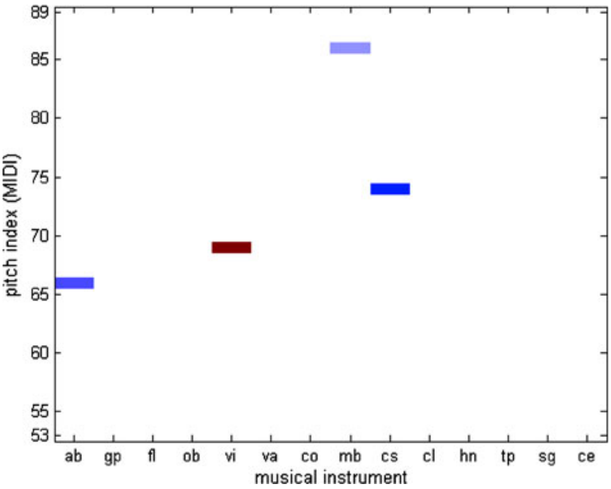

Fig. 3 A diagrammatic presentation of the weight matrix $\boldsymbol{\omega}$

Table 1 Description of musical instrument categories

Index	Instrument	Code	Index	Instrument	Code
1	Acoustic bass	ab	8	Marimba	mb
2	Grand piano	gp	9	Celesta	cs
3	Flute	fl	10	Clarinet	cl
4	Oboe	ob	11	French horn	hn
5	Violin	vi	12	Trumpet	tp
6	Viola	va	13	Steel guitar	sg
7	Church organ	co	14	Cello	ce

3 Experimental results

The experimental tests can be divided into three stages. The first stage aimed to measure the specific performance of our method with the assumption that all note onsets are accurate. The second stage aimed to assess the effects of onset misidentification in the process of instrument identification, pitch estimation and instrument-pitch identification. The third stage aimed to assess the performance of the entire system including using the note onset algorithm to locate onsets. In addition, we provide empirical comparisons to the methods proposed by Grindlay and Ellis (2010) and Kitahara et al. (2007).

3.1 Data

As our method has the capability of pitch estimation, to evaluate and quantify the performance of pitch estimation, we needed a set of polyphonic music compositions with an accurate MIDI reference. For this we used a dataset synthesized by Yamaha XG using the MIDI synthesis software, Cakewalk Pro Audio 9. The note models $P(f, t| p, m)$ were obtained from solo recordings of individual instruments. We used a set of 14 instruments (see Table 1) to derive our models, where each instrument had 37 pitches (F3–F6).

The test data were based on the chorale: "Aus meines Herzens Grunde", composed by Bach which contains four tracks. Without regard to the rationality of the combination of instruments, we arbitrarily chose two distinct instruments from the predefined 14 instruments to generate a mixture duet recording using a sample rate of 44.1 kHz. The duet recordings were produced by Cakewalk Pro Audio 9 in which the two selected instruments were played according to the MIDI ground truth of tracks 2 and 3. There were 91 synthetic duet recordings which were the mixtures of two instruments considering all possible combinations and the instruments all played one pitch at the same time. Only those simultaneous mixture notes within the natural pitch range of each instrument were used in the test. Among the tests, fewer mixture notes belonged to the acoustic bass than to the other instruments, because its pitch range was located in the lower section of the entire range of instruments, i.e. it had a small intersection with the other instruments in the given music score. A total of 3,129 duet notes were tested.

Similarly, the trios were produced according to the MIDI ground truth of the last three tracks, and the quartets were produced according to that of the entire four tracks. To obtain several simultaneous mixture notes as test samples, we allocated track 4, which consists mainly of lower pitches, to instruments with a lower pitch range, for instance the acoustic bass, cello and French horn; and track 1 consisting of higher pitches above G5 to instruments with a higher pitch range, for instance the grand piano, church organ, marimba and celesta.

For each test note, the durations of the fixed maximum frame length limit $\mathrm{T}_{\max} = 15$ (about 0.74 s) were ensured by cubic interpolation. The parameter α was set to 1.08 experimentally, while $\beta_1, \beta_2 \in [0.04, 0.14]$, and $\beta_3 \in [0.01, 0.3]$. In the process of salience measurement, we varied $\{\beta_i | i = 1, 2, 3\}$ to obtain the best results. The results in the following report were the selected best results from all the tests with diverse parameters.

3.2 Performance of the proposed method

This part of the tests aimed to verify which type of T-F representation was suitable, and synchronously measured the specific performance of our method in instrument identification, pitch estimation and instrument-pitch identification. In this subsection, the onset positions are assumed to be known, and all test samples are duet recordings.

Preliminary experiments were conducted to validate the design choices of the T-F representations introduced in Section 2.1. For simplicity, the T-F representation after the STFT, i.e. with linear and uniformly-spaced frequency, is hereinafter called the STFT, the T-F representation based on the semitone scale is called the semitone, and the T-F representation based on the ERB scale is called the ERB. The STFT had a cut-off frequency of 11 kHz, substantially reducing the frequency dimension. In spite of this, the STFT had the largest amount of data with 1,023 dimensions while the semitone had 96 dimensions and the ERB had 250 dimensions. The performance summaries are shown in Table 2 in which the results are the weighted average of all the instruments according to the corresponding number of samples:

$$\widetilde{F} = \frac{\sum_{m=1}^{M} F_m N_m}{\sum_{m=1}^{M} N_m} \tag{21}$$

F_m and N_m are the test results and the number of samples of the instrument m. Note that the test results in this paper were all quantified for each test recording in terms of the F-measure F (Vincent et al. 2010). As can be seen, the three T-F representations have very nearly the same performance in pitch estimation, however in the case of instrument identification and instrument-pitch identification the STFT performed the worst while the semitone and the ERB performed much better relatively. In other words, the frequency scale had little influence in pitch estimation but comparatively

Table 2 Summary F-measure of three T-F representations for the duet

T-F representation	STFT	Semitone	ERB
Instrument-pitch identification (%)	57.69%	87.16%	85.36%
Instrument identification (%)	59.90%	87.09%	84.73%
Pitch estimation (%)	97.38%	97.80%	97.30%

Note that the F-measure values are average results over all test data that included the same instrument greater influence in instrument identification and the resulting instrument-pitch identification. This confirmed that the log frequency scale (the semitone scale and the ERB scale are both log frequency scales), which is a good match with human hearing, can properly reveal the envelope of each harmonic (along the time axis) and the amplitude relationship between adjacent harmonics (along the frequency axis).

Overall, the best results for the proposed algorithm were obtained when the T-F representation used the semitone scale, and furthermore, the amount of data and computation required was the least among the three frequency scales. Consequently, the semitone scale was the preferred option for performing the following tests.

A full illustration of the accuracy of the identification parameters is shown as a function of the type of instrument in Fig. 4, in which a specific instrument in the horizontal ordinate indicates the mixture recordings containing this type of instrument and its results are an average of the tests on all these mixture recordings. Consequently, the test results of the mixture recording containing the acoustic piano and violin are counted twice in the terms of 'ap' and 'vi'. The analysis of Fig. 4 revealed some important information. First, the performance of pitch estimation was largely irrelevant to the type of instrument while the performance of instrument identification and instrument-pitch identification was directly relevant to the type of instrument. In general, the instrument that performed well for instrument identification also performed well for instrument-pitch identification. Moreover, those instruments with clean and distinct partials usually did well in instrument identification and instrument-pitch identification.

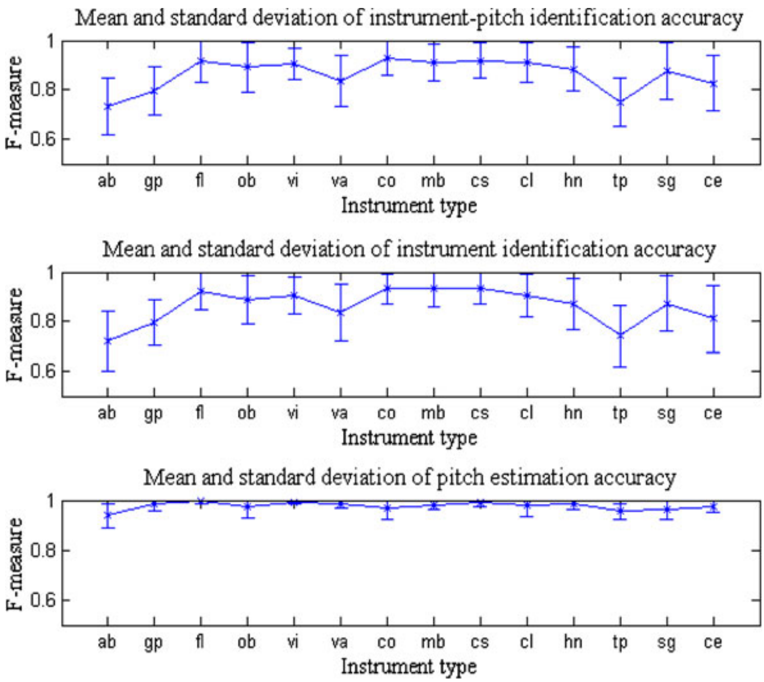

Fig. 4 Means and standard deviations of the accuracies of three tasks for duet recordings when the note onsets are known. The tests are based on the semitone scale T-F representation

For instance, the flute, church organ, marimba and celesta gave much better results than the acoustic bass, grand piano and trumpet.

3.3 Effects of onset misidentifications

Taking a duet recording as an example, this section analyses the effects of onset misplacements given the assumption that the offsets of notes were correct. The onset misplacement was divided into two types:

1. *The misidentified onsets are located after their actual positions*

In the case of pitch estimation, this type of misidentification had little effect on the performance of pitch estimation for almost all types of instruments, because even with a percussive instrument which generally has a quick-drop decay stage, there may be a faint partial (it is generally a fundamental frequency) with a long sustained stage in which the characteristics of the actual pitch still stand out. However, in the case of instrument identification and instrument-pitch identification, the performance of the mixture recordings containing such instruments, whose notes have a dominating sustained stage (e.g. flute and clarinet), gradually falls as the error in the onset increases, because the envelopes of the partials are approximately the same throughout their duration. While the performance of the mixture recordings containing percussive instruments, whose notes have a rapid notable attack and a tiny or scarcely any sustain (e.g., grand piano and marimba), falls rapidly, because the main content of the note, which is usually near the actual onset, may be lost. Taking all the instruments as a whole, the accuracy of both instrument identification and instrument-pitch identification falls slightly when the forward error of onset misplacement is smaller than 10 %, and slowly when it is larger than 10 %.

2. *The misidentified onsets are located before their actual positions*

In this case, the mixture signal as the input of proposed system includes not only an entire note but also part of a new note. This may confuse the frequency positions of the partials of the right note which would influence the performance of pitch estimation. Meanwhile, even if the pitches of the spurious segment are the same as those of the actual segment, the envelopes of partials vary too much and are hard to match to the right instrument models. Thus the severity of the effects will be directly linked to the backward error in pitch estimation, instrument identification and instrument-pitch identification. In particular, for percussive instruments, the accuracy of instrument identification and of instrument-pitch identification falls faster than for sustaining instruments.

Figure 5 summarizes the effects of both backward and forward onset misplacements (given as percentages of the actual frame length). The results were the average for all types of instruments, and the values shown in Fig. 5 represent the relative accuracy given by A_e/A_i, where A_e is the weighted-average accuracy of our proposed method when the onsets are misidentified, and A_i is the weighted-average accuracy of our proposed method when the onsets are in their correct positions. As can be seen, when forward onset misplacements existed in the test samples, the method still performed well in pitch estimation and the performance of the system fell slowly with forward error in instrument identification and instrument-pitch identification.

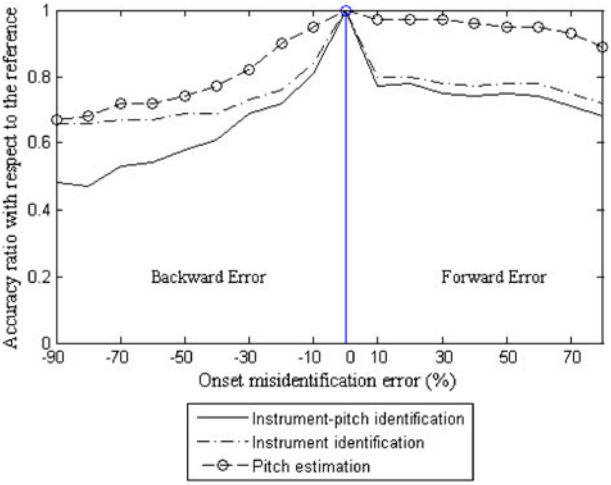

Fig. 5 Effect of the backward and forward onset misplacements

However, when backward onset misplacements existed in the test samples, the accuracy for all three cases fell, especially for instrument-pitch identification.

On the whole, the errors smaller than 10 % of the frame length (including backward and forward errors) had less impact on the accuracy for each case because the characteristics of partials were only slightly altered. In addition, the backward errors had a greater effect on the performance of the system than the forward errors. Finally, the onset misidentification had less effect on the performance for pitch estimation than for instrument identification and instrument-pitch identification.

3.4 Performance of the entire system

The results presented in this subsection were obtained using the entire system where the note onsets were located by an onset detection algorithm proposed in our early research (Hu and Liu 2011). First, taking a duet as an example, our entire system was compared with two algorithms; the PET system (Grindlay and Ellis 2010) and Instrogram (Kitahara et al. 2007). Our PMMD system and the PET system were all tested based on the note level whereas the Instrogram system was based on the frame level. Table 3 shows the contrasting results in that the PET system showed almost no ability for instrument identification and the Instrogram system performed poorly in allocating the right pitches to respective instruments. From these comparisons, we can see that our PMMD system outperformed the two systems in pitch estimation, instrument identification and instrument-pitch identification.

Table 3 Performance comparisons between our entire system and others

Algorithm	Proposed PMMD	PET algorithm	Instrogram
Instrument-pitch identification (%)	87.16%	12.79%	9.78%
Instrument identification (%)	87.09%		20.41%
Pitch estimation (%)	97.80%	84.33%	88.71%

Our proposed entire PMMD algorithm was based on the semitone scale. The PMMD and PET were tested based on the note level whereas the Instrogram was based on the frame level

To further verify the performance of our entire system, the trios and the quartets were also used in the. The performance of the system when the onsets were located using the onset detection algorithm was also in contrast to its performance when the onsets were actually correct. With the definition that if the detected onsets and target onsets are within an ±23 ms time window (one frame), the detected onsets seem accurate.

The average accuracy of onset detection for duets, trios and quartets were 90.31, 97.54, and 98.1 %. Note that for all the test audio recordings played by two to four instruments, each instrument generated one pitch at a specific time. All tests were based on the note level with the T-F representation on the semitone scale. The results are shown in Table 4 in terms of the number of instruments in the test recordings. In this table, the *number of instruments* row shows the total number of simultaneous instruments presented in the mixture recording. As can be seen, the results degraded significantly as the number of instruments increased. This is because more instruments imply that there will be a more sophisticated combination of partials, resulting in more difficult decomposition of the mixture model. In contrast to the results when the onsets are correct, for each type of mixture recording, the results all deteriorated to various degrees when the onsets were located using the note onset detection algorithm. In particular, the performance of the entire system for duets deteriorated most severely, which is related to our test dataset. Most duet recordings do not contain percussive instruments whereas the great majority of trio recordings and almost all quarter recordings do. Thus the note onsets detected for duet recording are more inaccurate than for quartet and trio recordings, as the note onset detection algorithm used in our system performed fairly well on percussive signals. In addition, according to our experimental data, we found that few errors existed on onset misidentification for percussive trio and quarter recordings, and even where they did exist the vast majority was within ±10 % of the frame length. For non-percussive duet recordings, more errors of onset misidentification existed than for percussive recordings, and they were generally backward errors. Therefore, in our entire system, for duet recordings, the performance of instrument identification and instrument-pitch identification reduced by about 12 and 17 % while the performance of pitch estimation reduced by about 7 %. These confirmed two points: first, the backward errors have a greater effect on the performance of the system; second, the onset misidentification had less effect on the performance of pitch estimation than on instrument identification and instrument-pitch identification. For percussive trio and quartet recordings, the performance for all three cases reduced slightly. This experimental consequence verified that the errors within ±10 % of the frame length

Table 4 Condensed comparison between the system employing the note onset detection tool and the system with completely correct note onsets

	Note onsets are correct (%)			Note onsets are located by onset detection algorithm (%)		
Number of instruments	2	3	4	2	3	4
Instrument-pitch identification	87.16	69.13	60.12	70.11	68.32	55.41
Instrument identification	87.09	73.79	66.58	75.38	73.22	62.49
Pitch estimation	97.80	93.54	87.30	91.24	92.77	86.01

had less impact on the accuracy of pitch estimation, instrument identification and instrument-pitch identification, as concluded in the previous subsection.

4 Conclusions

This paper presented a new method for estimating pitch, identifying instruments and simultaneously assigning each pitch to its source instruments in monaural polyphonic audio recordings containing multiple instruments. The method uses a probabilistic mixture model decomposition approach to determine the type of instrument and pitches in the mixture recordings, and allocates these pitches to their corresponding instruments. The method did not need to know the number of pitches and instruments in advance. The method has comparable or possibly even superior performance to the human ear. A possible shortcoming of the proposed algorithm is its dependency on the tools for note onset detection and its reliance on only one sample when generating each note model. Tests have shown that, although the errors caused by onset detection tools indeed propagate through the whole system, their overall effect is actually tolerable. In contrast to the other two methods, our entire system was still superior, especially for instrument identification and instrument-pitch identification.

We tested only on synthetic mixture signal, as it is difficult to evaluate and quantify the performance of real recordings, from which one can hardly obtain the accuracy and reliable reference scores, especially the pitch reference scores. However, in fact, the real recordings are much more complicated and diversified because of diverse players and instruments etc. Thus it is necessary to assess the performance of our system against real recordings in the future work.

Future work will also concentrate on reducing the dependency on the onset detection algorithm and extending the models generated using increased training samples. The results presented here can be extended and expanded in a number of applications. We expect that the algorithm proposed here will enhance applicability and effectiveness in a number of music applications.

Acknowledgements The authors would like to thank E. Vincent for sharing the ERB code of his algorithm, and J.C. Brown for sharing the CQT code of his algorithm. This work is supported by the National Natural Scientific Foundation of China Project No.61173110 and Key Projects in the National Science & Technology Pillar Program 2011BAK08B02.

References

Barbedo, J. G. A., & Tzanetakis, G. (2011). Musical instrument classification using individual partials. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(1), 111–122.

Bay, M., & Beauchamp, J. (2006). Harmonic source separation using prestored spectra. In *Indep. Compon. Anal. and Blind Signal Separ.* (pp. 561–568).

Bertin, N., Badeau, R., Vincent, E. (2009). Fast Bayesian NMF algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription. In *IEEE Workshop Appl. Signal Process. Audio Acoust.* (pp. 29–32). NY, USA: New Paltz.

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute, 4*, 126.

Brown, J. C. (1991). Calculation of a constant Q spectral transform (Vol. 89, Vol. 1): *Vision and modeling group, media laboratory*, Massachusetts Institute of Technology.

Burred, J.J., Robel, A., Sikora, T. (2010). Dynamic spectral envelope modeling for timbre analysis of musical instrument sounds. *Audio, Speech, and Language Processing, IEEE Transactions on, 18*(3), 663–674.

Dessein, A., Cont, A., Lemaitre, G. (2010). Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *Int. soc. for music inf. retrieval conf., Utrecht, Netherlands*.

Dziubinski, M., Dalka, P., Kostek, B. (2005). Estimation of musical sound separation algorithm effectiveness employing neural networks. *Journal of Intelligent Information Systems, 24*(2), 133–157.

Essid, S., Richard, G., David, B. (2006). Musical instrument recognition by pairwise classification strategies. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(4), 1401–1412.

Goto, M. (2004). A predominant-F0 estimation method for polyphonic musical audio signals. In *Proc. int. cong. on acoustics, ICA* (pp. 1085–1088).

Grindlay, G., & Ellis, D.P.W. (2010). A probabilistic subspace model for multi-instrument polyphonic transcription. In *Int. soc. for music inf. retrieval conf., Utrecht, Netherlands* (pp. 21–26).

Heittola, T., Klapuri, A., Virtanen, T. (2009). Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *Int. soc. for music inf. retrieval conf., Kobe, Japan* (pp. 327–332).

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *ACM proceedings of twenty-second annual int. SIGIR conf* (pp. 50–57). New York: ACM.

Hu, Y., & Liu, G. (2011). Dynamic characteristics of musical note for musical instrument classification. In *IEEE int. conf. on signal processing, communications and computing* (pp. 1–6). Xi'an, China: IEEE.

Jiang, W., Wieczorkowska, A., & Raś, Z. (2009). Music instrument estimation in polyphonic sound based on short-term spectrum match. *Foundations of Computational Intelligence, 2*, 259–273.

Joder, C., Essid, S., Richard, G. (2009). Temporal integration for audio classification with application to musical instrument classification. *Audio, Speech, and Language Processing, IEEE Transactions on, 17*(1), 174–186.

Kameoka, H., Nishimoto, T., Sagayama, S. (2007). A multipitch analyzer based on harmonic temporal structured clustering. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(3), 982–994.

Kitahara, T., Goto, M., Komatani, K., Ogata, T., Okuno, H.G. (2007). Instrogram: probabilistic representation of instrument existence for polyphonic music. *Information and Media Technologies, 2*(1), 279–291.

Kostek, B. (2004). Musical instrument classification and duet analysis employing music information retrieval techniques. *Proceedings of the IEEE, 92*(4), 712–729.

Kursa, M., Rudnicki, W., Wieczorkowska, A., Kubera, E., Kubik-Komar, A. (2009). Musical instruments in random forest. *Foundations of Intelligent Systems*, 281–290.

Li, Y., Woodruff, J., Wang, D.L. (2009). Monaural musical sound separation based on pitch and common amplitude modulation. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(7), 1361–1371.

Loughran, R., Walker, J., O'Neill, M., O'Farrell, M. (2008). The use of mel-frequency cepstral coefficients in musical instrument identification. In *Proc. of the international computer music conference (ICMC), SARC, Belfast, N. Ireland*.

Rao, P., & Shandilya, S. (2004). On the detection of melodic pitch in a percussive background. *Journal of Audio Engineering Soc., 52*(4), 378–391.

Shashanka, M., Raj, B., Smaragdis, P. (2008). Probabilistic latent variable models as nonnegative factorizations. *Computational Intelligence and Neuroscience, 2008*, 947438.

Smaragdis, P., Raj, B., Shashanka, M. (2006). A probabilistic latent variable model for acoustic modeling. In *Advances in Models for Acoustic Processing, NIPS* (Vol. 146).

Vincent, E., Bertin, N., Badeau, R. (2010). Adaptive harmonic spectral decomposition for multiple pitch estimation. *Audio, Speech, and Language Processing, IEEE Transactions on, 18*(3), 528–537.

Wieczorkowska, A.A., & Kubera, E. (2010). Identification of a dominating instrument in polytimbral same-pitch mixes using SVM classifiers with non-linear kernel. *Journal of Intelligent Information Systems, 34*(3), 275–303.

Wu, J., Vincent, E., Raczynski, S., Nishimoto, T., Ono, N., Sagayama, S. (2011). Polyphonic pitch estimation and instrument identification by joint modeling of sustained and attack sounds. *IEEE Journal of Selected Topics in Signal Processing, 5*(6), 1124–1132.