Music Signal Decomposition Based On Sequential Identification and Subtraction of Components

Štepán Albrecht[1], Václav Matoušek[2]

[1] *University of West Bohemia in Pilsen, Univerzitní 8, 306 14 Plzen, Czech Republic, Email: albrs@kiv.zcu.cz*

[2] *University of West Bohemia in Pilsen, Univerzitní 8, 306 14 Plzen, Czech Republic, Email: matousek@kiv.zcu.cz*

Abstract

In recent years there has been much effort towards automatic content-based music information retrieval. Such systems first analyze the audio content to obtain a MIDI-like representation, then pattern matching techniques are applied in order to identify what one desires. The task is to find a feasible MIDI-like representation. In this paper, the idea of discovering the MIDI-like representation follows the operation of music composing software tools called "trackers" – a music is done by putting the sounds into the tracks. Our concept of music/audio signal decomposition (ASD) is a reverse process. Bayesian statistical methods are applied.

Introduction

There is an original observed audio signal (a song) $\mathbf{Y}$ being a composition of anything – from the tone A of the piano to the drum loop in a popular song (the audio component). Then we are given a *set* of audio components and we expect that it is possible to combine the observed signal from the audio components in the *set* or their parts. We can examine the two tasks:

1. Identify the components on the positions in time, or their possible transformations.
2. To investigate the conditions while the *set* of samples is large enough w.r.t. the transformations and the observed signal may be recovered.

The first is a starting point to the ASD concept, therefore this paper is aimed at this, specifically, on identification of the components or their parts and finding their start times. Every component consists of consecutive frames. A frame is a vector of windowed Discrete Fourier Transform gained from the audio signal of a short duration (0.05-0.1s). All frames of all components build matrix $\mathbf{X}$.

Selection of Algorithms

This Music decomposition is a complex problem and deterministic approaches do not allow to incorporate all useful information (heuristics) into the solution of this, hence we are made to search for a large space to find the optimal or sub-optimal parameters. Bayesian estimation allows this, but exact parameter estimation is not usually possible due to the intractable formulas. However, there exists a Monte Carlo (MC) methods capable to approximate them. The approximation is feasible at the cost of generating many samples, which is not as much time consuming for such complex problem as the searching for in the large parameter space. Author's former idea was a Matching-Pursuit-like iteration algorithm [2]. It consisted of the two steps – identification of a predominant component and its subtraction. However, first tests on identification proved that the local detection of components (i.e., without considering other components that could be present at the same time) was not reliable. When the simultaneous components were considered, there was a huge amount of computation load to get the first predominant candidates, moreover we did not regard all knowledge we could. That is why we decided to process its identification part by Sequential Monte Carlo (SMC). We guess that knowledge information incorporated into SMC will allow us to abandon the subtraction step. The algorithm of the ASD introduced here is currently being tested.

Monte Carlo in General

Bayes's theory provides a general framework for statistical inference. The main concept is that of posterior distributions, which result from both the likelihood and the prior. MC methods approximate the *minimum mean square error* (MMSE) [2]

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MMSE}} = \int_{\theta} \boldsymbol{\theta} \mathrm{p}(\boldsymbol{\theta}|X) d\boldsymbol{\theta} \approx \frac{1}{M} \sum_{i=1}^{M} \tilde{\boldsymbol{\theta}}^{(i)}. \qquad (1)$$

We can see that as soon as the samples $\tilde{\boldsymbol{\theta}}^{(i)}$ are obtained by sampling[1] from the posterior $\tilde{\boldsymbol{\theta}}^{(i)} \sim \mathrm{p}(\boldsymbol{\theta}|X)$, the problem is solved. That implies that the problem solution resides in computation of the posterior distribution which enables sampling. Several technics has been developed for random variable generation from any pdf. We deal with the Sequential Importance Sampling (SIS).

Sequential Importance Sampling

Importance sampling (IS) is a trick how to tackle the problem of being unable to sample from the posterior. It is drawn from an another distribution $\mathrm{q}(\boldsymbol{\theta}|X)$ called the *importance*. The idea is represented by equation.

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MMSE}} = \int_{\theta} \boldsymbol{\theta} \mathrm{p}(\boldsymbol{\theta}|X) d\boldsymbol{\theta} = \int_{\theta} \boldsymbol{\theta} . \frac{\mathrm{p}(\boldsymbol{\theta}|X)}{\mathrm{q}(\boldsymbol{\theta}|X)} . \mathrm{q}(\boldsymbol{\theta}|X) d\boldsymbol{\theta} \qquad (2)$$

[1] i.e. generating

Thus, the MC samples are weighted during averaging

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MMSE}} \approx \frac{1}{M}\sum_{i=1}^{M}\tilde{\omega}^{(i)}\tilde{\boldsymbol{\theta}}^{(i)}, \quad \tilde{\omega}^{(i)} = \frac{\mathrm{p}(\boldsymbol{\theta}^{(i)}|X)}{\mathrm{q}(\boldsymbol{\theta}^{(i)}|X)}, \quad \mathrm{p,q} \neq 0. \tag{3}$$

Now, let us consider the non-linear state space model, satisfying the Hidden Markov Model assumptions. Then the weight computation is governed by the recursive [3]

$$\tilde{\omega}_t^{(i)} = \frac{\mathrm{p}(\boldsymbol{\theta}_{0:t}^{(i)}|X_{1:t})}{\mathrm{q}(\boldsymbol{\theta}_{0:t}^{(i)}|X_{1:t})} = \omega_{t-1}^{(i)}\frac{\mathrm{p}(X_t|\boldsymbol{\theta}_t^{(i)})\mathrm{p}(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)})}{\mathrm{q}(\boldsymbol{\theta}_t^{(i)}|X_t,\boldsymbol{\theta}_{t-1}^{(i)})}. \tag{4}$$

Sequential generation of samples and calculation of weights is called Sequential Importance Sampling[2]. Getting the optimal $\mathrm{q}(\boldsymbol{\theta}_t^{(i)}|X_t,\boldsymbol{\theta}_{t-1}^{(i)})$ is done preferably analytically, if it is not possible (our case), we try to approximate it with an sub-optimal importance (that is a heuristic) reflecting observed data X_t and past hidden data $\boldsymbol{\theta}_{t-1}$.

Model

We consider a linear sequential model of superposition for every processed observed signal frame $\mathbf{y}_t$, for the matrix of all N_{all} frames $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|\ldots|\mathbf{x}_{N_{\mathrm{all}}}]$ and the vector of frame presences $\mathbf{s}_t$. We have

$$\mathbf{y}_t \approx \mathbf{X}\mathbf{s}_t. \tag{5}$$

The vector of presences $\mathbf{s}_t$ contains either one or zero on position i w.r.t. the presence or non-presence frame x_i. The vector of one-positions in $\mathbf{s}_t$ is denoted by $\mathbf{n}_t$. Length of $\mathbf{n}_t$ is represented by N_t and the maximal polyphony is limited to $N_{\max}$. The observed signal $\mathbf{Y} = [\mathbf{y}_1|\mathbf{y}_2|\ldots|\mathbf{y}_T]$ is frame-wise processed by the SMC and the components or their parts are sequentially identified.

Our observed data at time t is the vector $\mathbf{y}_t$. Hidden parameters $\boldsymbol{\theta}_t$ to discover consists of partly the vector of indices $\mathbf{n}_t$ of frames $\mathbf{x}_i, i = 1\ldots N_{\mathrm{all}}$, partly the length N_t of $\mathbf{n}_t$ and partly the accompanying hyperparameters[3] $\boldsymbol{\psi}_t$. We would like to sample from the posterior

$$\begin{aligned}\mathrm{p}(\mathbf{n}_t,\boldsymbol{\psi}_t|\mathbf{y}_t,N_t) \quad &\propto \quad \mathrm{p}(\mathbf{y}_t|\boldsymbol{\psi}_t,\mathbf{n}_t).\mathrm{p}(\mathbf{n}_t|\mathbf{n}_{t-1},N_t) \\ &\times \quad \mathrm{p}(\boldsymbol{\psi}_t|\boldsymbol{\psi}_{t-1}).\mathrm{p}(N_t|N_{t-1}),\end{aligned} \tag{6}$$

where $\mathrm{p}(\mathbf{y}_t|\boldsymbol{\psi}_t,\mathbf{n}_t) = \mathcal{N}(\mathbf{y}_t - \mathbf{X}_t.\mathbf{s}_t; 0, \boldsymbol{\Sigma}_1)$ is the likelihood of $\mathbf{y}_t$ being observed given the hidden parameters. Heuristics $\mathrm{p}(N_t|N_{t-1})$ says which polyphony is more probable given the polyphony at $t-1$. It is given by a transitional table. Furthermore, it ensures, that the polyphony will not exceed $N_{\max}$. $\mathrm{p}(\mathbf{n}_t|\mathbf{n}_{t-1},N_t)$ is a prior distribution for the indices of simultaneous frames which is obtained from the transition table H more laboriously (see table 1). The prior reflects a higher probability of generating a frame index which is a continuation of the frame present at time $t-1$. $\mathrm{p}(\boldsymbol{\psi}_t|\boldsymbol{\psi}_{t-1}) = \mathcal{N}(\boldsymbol{\psi}_t; \boldsymbol{\psi}_{t-1}, \boldsymbol{\Sigma}_2)$ is a prior (also a transitional) for the hyperparameters. Calculation of the posterior, which the hidden parameters could be generated from, is not possible since the distribution $\mathrm{p}(\mathbf{n}_t|\mathbf{n}_{t-1},N_t)$ is non-linear, non-gaussian and

[2]also called Particle Filtering

[3]e.g., variance of the likelihood

	1	2	3	4
1		1		
2			1	
3				0.8
4	0.4			

Tabulka 1: Example of the table H of transitions $\mathrm{p}(n_t|n_{t-1})$ between 4 frames. Rows – previous frames, columns - upcoming frames (n is not a vector now).

even the number of items $\mathbf{n}_t$ varies. Therefore we follow the way of IS – the hidden parameters $\boldsymbol{\theta}_t = [N_t, \mathbf{n}_t, \boldsymbol{\psi}_t]$ are obtained by generating the MC samples from their importance distributions. $N_t, \boldsymbol{\psi}_t$ are generated from their priors (see the algorithm). However, importance distribution sampling $\mathbf{n}_t \sim \mathrm{p}(\mathbf{n}_t|\mathbf{y}_t,\mathbf{n}_{t-1},N_t)$ is formulated in Bayesian manner (heuristics from the data $\mathbf{y}_t$ times the prior), thus enabling to generate samples of high probability in the posterior distribution.

Algorithm

Initializing: $\Rightarrow \tilde{\omega}_0^{(i)} = 1, \tilde{\boldsymbol{\theta}}_0^{(i)} \sim \mathrm{q}(\boldsymbol{\theta}_0)$.
Iterations:
$\Rightarrow$ For i=1 to M: $\tilde{N}_t^{(i)} \sim \mathrm{p}(N_t|N_{t-1}), \tilde{\mathbf{n}}_t{}^{(i)} \sim \mathrm{p}(\mathbf{n}_t|\mathbf{y}_t,\mathbf{n}_{t-1},N_t), \tilde{\boldsymbol{\psi}}_t{}^{(i)} \sim \mathrm{p}(\boldsymbol{\psi}_t|\boldsymbol{\psi}_{t-1})$;
$\Rightarrow$For i=1 to M calculate the reduced formula of weights $\tilde{\omega}_t^{(i)} \propto \omega_{t-1}^{(i)}\frac{\mathrm{p}(\mathbf{y}_t|\boldsymbol{\psi}_t,\mathbf{n}_t).\mathrm{p}(\mathbf{n}_t|\mathbf{n}_{t-1},N_t)}{\mathrm{p}(\mathbf{n}_t|\mathbf{y}_t,\mathbf{n}_{t-1},N_t)}$
$\Rightarrow$ Weighted average hidden p. calculation (see 3).
$\Rightarrow$ Resampling step – due to possible weight degeneration, i.e. few weights are of large values compared to the others [2],[3].

Testing

The tests have not been done yet. We propose the following approach: a part of first component and another part of the second are superposed and reverberated, thus we get the testing $\mathbf{Y}$.

Conclusions, Future Work

We introduced the concept of ASD. We explained the reasons which leaded us to apply the SMC methods. We described the model governed by the SMC and the algorithm. We informed about the testing of this approach. In the end we notice that if the algorithm is to be applied for real data, the distributions and heuristics dealing with the observed signal have to operate on mid-level representation of $\mathbf{Y}$ [1], not on mere DFT or Mel frequency cepstral coefficients.

Reference

[1] "Bello, J., Pickens J.: A Robust Mid-level Representation for Harmonic Content in Music Signals, ISMIR Proceedings, 2005",

[2] "Klapuri, A., Davy M., "Signal Processing Methods For Music Transcription", Springer, 2006",

[3] "Doucet, A., Godsill, S., Andrieu, C.: On Sequential Monte Carlo Sampling methods for Bayesian filtering, Statistics and Compution, 2000".