# On the Inverse Music Sequencer Operation – Detection of Music Components from Wave Table in Complex Music Signal

Š. Albrecht

*University of West Bohemia, Czech Republic, Email: albrs@kiv.zcu.cz*

## Introduction

The goal of this paper is to design of a novel entire system for automatic music transcription (AMT). By the complete AMT we mean a process of resolving the pitch, loudness, timing and instrument of all sound events in an input audio music signal [4]. Complete music signal transcription is not theoretically possible [4], therefore the entire systems are a subject to a conception of what are the expectations, e.g.,

- audio scene analysis (Kashino et. al.) [7].
  - aims at extracting entities like notes, chords, beats, rhythm from an audio signal.
  - operates on psychophysical findings regarding the acoustic "cues" that humans use to assign the spectral components to their respective sources.
  - utilizes a set of internal models involving information of chord, note, frequency component progression.
- music scene description (Goto et. al.) [6]
  - the aim is to obtain descriptions that are intuitively meaningful to an untrained listener.
  - the purpose is not to extract every musical note.
  - includes the analysis of melody, bass lines, metrical structure, rhythm and chorus and phrase repetition.

## Conception

Music sequencers use the library of sounds (sound components) to compose or play music. This conception follows the inverse working of music sequencers (see Fig. 1), that is, given the library of sounds the input complex music signal is passed into the decomposition process in which in sound events are found. Time, ID and a modification type of a library sound are considered as the sound events. Component modification(s) can reduce the library size. Amplitude altering, sound component truncation, pitch shifting can be considered as such modification. It has to be noted that arbitrary sounds in the library are allowed in general. Since the library sounds can be either automatically or manually described with MIDI information (i.e., instrument type, notes contained, etc.) the entire MIDI information of the music can be retrieved.

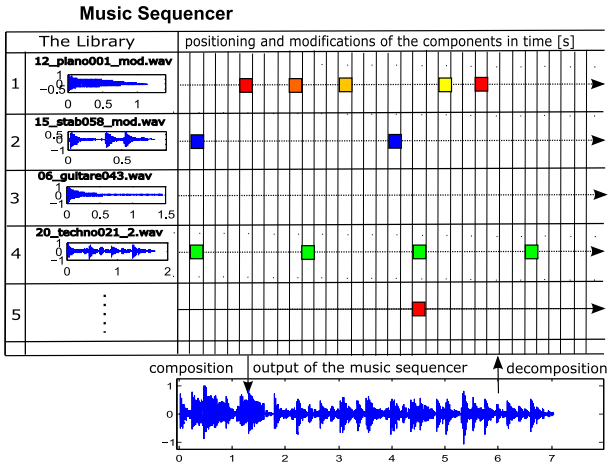


**Figure 1:** Library sound modification type depicted by a square of a distinct color.

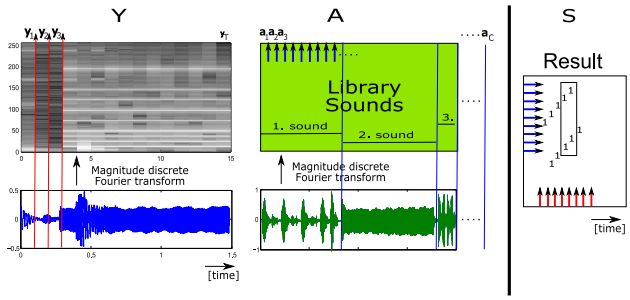


**Figure 2:** Example of an input signal composed of two overlapped library sounds. Vectors $\mathbf{s}_t, t = 1 \ldots T$ form the matrix $\mathbf{S}$. Blanks in $\mathbf{S}$ stand for zeros.

## Music Signal Model – Structural Description

Signal composition equation is defined as

$$\mathbf{y}_t \approx \mathbf{A} \cdot \mathrm{diag}(\mathbf{s}_t)\beta_t$$

where $\mathbf{y}_t$ is a magnitude DFT frame (segment) of an input signal at time $t$, $\mathbf{A}$ the library representation, magnitude DFT frame (segment) $c$ of a library sound is represented by a vector $\mathbf{a}_c$; $\mathbf{A} = [\mathbf{a}_1|\mathbf{a}_2|..|\mathbf{a}_C]$, $\mathrm{diag}(\mathbf{s}_t)$ represents a matrix, which diagonal is $\mathbf{s}_t$ vector of elements one, zero (frame presence or non-presence). Polyphony number $N_t$ denotes number of ones at $t$, $N_t < N_{\max}$, where $N_{\max}$ is a user-specified parameter. Amplitude is computed as $\beta_t$ $\hat{\beta}_t = \mathrm{argmin}_{\beta_t} ||\mathbf{y}_t - \mathbf{A} \cdot \mathrm{diag}(\mathbf{s}_t)\beta_t||$. The structural description is depicted in Fig. 2. Component truncation is considered as the modification type and its parameters are discovered by joining the component frames.

The description can be interpreted as a *Markov model*

with a state defined as $\mathbf{x}_t = [N_t, \mathbf{s}_t]$ evolving according to

$$\begin{aligned} \mathbf{y} &= g(\mathbf{x}_t) + \mathbf{w}_t \qquad \text{(observation equation)}, \\ \mathbf{x}_t &= f(\mathbf{x}_{t-1}, \mathbf{v}_t) \qquad \text{(transitional equation)} \end{aligned}$$

# Music Signal Model – Probabilistic Description

Observation equation represents the likelihood $\mathrm{p}(\mathbf{y}_t|\mathbf{s}_t, N_t) = \mathcal{N}(\mathbf{y}_t; \mathbf{A} \cdot \mathbf{S}_t\beta_t, \Sigma)$, where $\mathbf{w}_t$ is a Gaussian noise with zero mean and covariance $\Sigma$. Transitional equation represents the state transitional probability density function (pdf) $p(\mathbf{x}_t|\mathbf{x}_{t-1})$. The transitional equation is defined for the inner state variables $\mathbf{s}_t, N_t$ separately; for the frame presence vector $\mathbf{s}_t$ the transitional distribution $\mathrm{p}(\mathbf{s}_t|\mathbf{s}_{t-1})$ is formed upon the table 1(b). Higher probabilities are assigned to transitions of successive frames. Transition distribution $\mathrm{p}(N_t|N_{t-1})$ is given by Markov model with transition matrix see the table 1(a). The aim is to allow $N_t$ to increase or decrease but the probability to keep $N_t$ constant must be prominent.

| From / To | 1 | 2 | 3 | 4 | ... |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1 | ... |
| 2 | 1 | 0 | 0 | 0 | ... |
| 3 | 0 | 1 | 0 | 0 | ... |
| 4 | 0 | 0 | 1 | 0 | ... |
| ... | ... | ... | ... | ... | ... |

(a)

| $N_{t-1}$ / $k_t$ | 0 | 1 | 2 |
|---|---|---|---|
| +2 | 1/10 | 0 | 0 |
| +1 | 2/10 | 1/9 | 0 |
| 0 | 7/10 | 7/9 | 7/10 |
| -1 | 0 | 1/9 | 2/10 |
| -2 | 0 | 0 | 1/10 |

(b)

**Table 1:** Tab. 1(b) represents the example of transition probabilities of 4 frames. In 1(a) is transitional distribution for number of simultaneous active frames $N_t$, which is given by $k_t \sim \mathrm{p}(k_t|N_{t-1})$ and by $N_t = N_{t-1} + k_t$. In the tables $N_{\max} = 2$ is considered.

Likelihood, state transition density and initial state distribution $\mathrm{p}_0(\mathbf{x}_0)$ define probabilistic description of discrete state space model.

# State Estimation and Bayesian Approach

The posterior distribution $\mathrm{p}(\mathbf{x}_t|\mathbf{y}_{1:t})$ given by Bayes's theorem can be expressed by the recursive formula

$$\mathrm{p}(\mathbf{x}_t|\mathbf{y}_{1:t}) = \mathrm{p}(\mathbf{y}_t|\mathbf{x}_t) \int \frac{\mathrm{p}(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\mathrm{p}(\mathbf{x}_t|\mathbf{x}_{t-1})}{\mathrm{p}(\mathbf{y}_t|\mathbf{y}_{1:t-1})} d\mathbf{x}_{t-1}$$

We are aimed at point estimate of the state $\mathbf{x}_t$. The estimation reflects a criterion (e.g, minimum-mean square error (MMSE) or maximum a posteriori (MAP)), the estimation is given by the posterior density – the integral must be tractable, but it is usually not.

*(Sequential) Monte Carlo* (SMC) methods are based on the fact that having enough samples reflecting some phenomenon, the distribution of the phenomenon can be approximated and point estimates can be obtained [3], [5]. The MMSE point estimate of $\mathbf{x}_t$ from samples $\mathbf{x}_t^{(i)}$ is then simply $\hat{\mathbf{x}}_t = \frac{1}{M}\sum_{i=1}^{M} \mathbf{x}_t^{(i)}$.

# Sequential Monte Carlo

In the previous section we encountered problem that the posterior is not usually tractable. The solution of the problem is to avoid this intractable calculations by sampling from another pdf q called *proposal*. The proposal should be as close as possible to the posterior. Discrepancy between the posterior and proposal is compensated by *weight*:

$$\omega_t^{(i)} = \frac{\mathrm{p}(\mathbf{x}_t^{(i)}|\mathbf{y}_{1:t})}{\mathrm{q}(\mathbf{x}_t^{(i)}|\mathbf{y}_{1:t})} \propto \frac{\mathrm{p}(\mathbf{y}_t|\mathbf{x}_t^{(i)})\mathrm{p}(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)})}{\mathrm{q}(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)}, \mathbf{y}_t)} .\omega_{t-1}.$$

When $\sum_{i=1}^{M} \omega_t^{(i)} = 1$, the MMSE point estimate is obtained as: $\hat{\mathbf{x}}_t = \sum_{i=1}^{M} \omega_t^{(i)} \mathbf{x}_t^{(i)}$. It is advantageous that the normalization term $\mathrm{p}(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ need not be computed since it is reduced in the weight formula fraction. The estimation algorithm [1], [2] is given as:

**Algorithm** *SMC Estimation Algorithm*:

- Every time step $t$
  - $M$ samples are generated from the importance distribution,
  - point estimate of the parameters $\mathbf{s}_t, N_t$ are calculated using the weights and samples,
  - resampling step [2].

# Experimental Setup and Testing

Proposal q deals with the inner state variables $\mathbf{s}_t, N_t$ independently, i.e., for $N_t$ is equivalent to the transitional pdf, for $\mathbf{s}_t$: Sampling mechanism and probability calculation is based on the similarity measure calculation MEASURE $= \frac{\mathbf{y}_t.\mathbf{a}_c}{||\mathbf{y}_t||.||\mathbf{a}_c||}$ and favoring successive frames of one component.

## Test #1

See Figure 3. Testing sound was created from two components of the library. Both were truncated and the truncation parameters were recorded. The sound components were overlapped so that the longer component started at the half of the shorter – they were summed resulting in a testing sound, thus, we had exact information about the truncations, the number of the components and their times.

The library **A** contained 21 sound components of arbitrary length, about 30 seconds of a complex music signal together (354 non-overlapping frames), the DFT frame

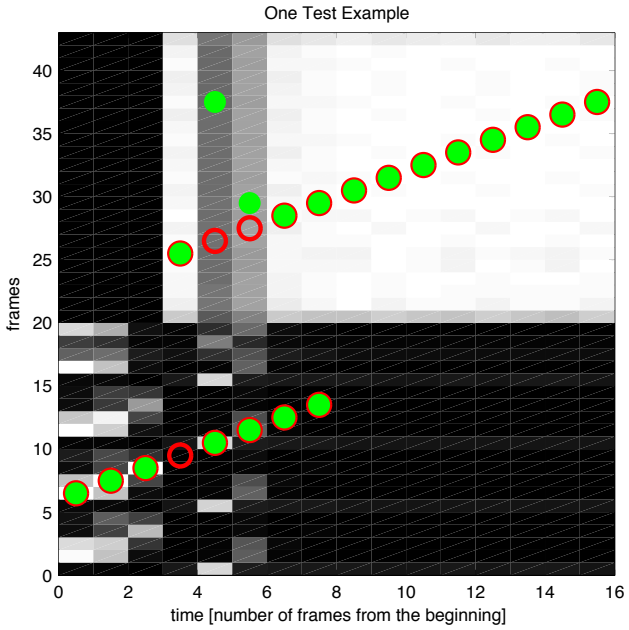


**Figure 3:** In the figure – representation of the hits regarding the similarity measure (black-white background), i.e., part of computation of the proposal distribution.

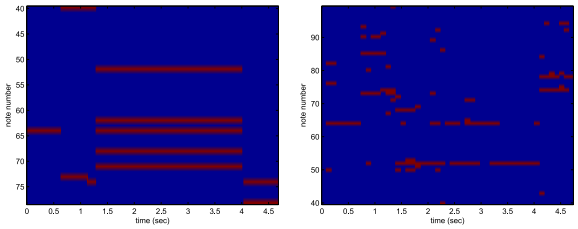


**Figure 4:** The piano transcription.

of length 4096 samples, signal sample rate at 44.1 kHz, silent frames $\mathbf{y}_t$ were treated individually. $M = 2000$ samples, $N_{\max} = 2$, $\Sigma$ had diagonal values of 0.5, while the amplitude of a sound wave was up to one.

### Test #2

Figure 4. Transcription of Chopin music excerpt recorded from a midi piano. The library contained one audio file with 64 piano tones synthesized with VST instruments, together about 55s, $M = 5000$, $N_{\max} = 4$, other settings the same as in the Test #1. MIDI transcription allowed since every frame associated with a MIDI note.

## Conclusion and Future Work

Results are not optimal yet since information from $t, t-1$ (Markovian) for $\mathbf{x}_t$ estimation does not contain much information about $\mathbf{x}_{t-1}, \mathbf{x}_{t-3}, \ldots$. The model should be further adjusted while Markovian property is not broken. More appropriate likelihood and similarity measure will be further considered.

## References

[1] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, 50(2):174–188, 2002.

[2] Z. Chen. Bayesian filtering: From kalman filters to particle filters, and beyond. *IEEE Transactions on Signal Processing*, 50(2).

[3] M. Davy and C. Dubois. A fast particle filtering approach to bayesian tonal music transcription.

[4] M. Davy and A. Klapuri, editors. *Signal Processing Methods For Music Transcription*. Springer, 2006.

[5] C. Dubois. Harmonic tracking using sequential monte carlo. In *in SSP*, 2005.

[6] M. Goto. Music scene description project: Toward audio-based real-time music understanding. In *4th International Conference on Music Information Retrieval*.

[7] K. Kashino and H. Tanaka. A sound source separation system with the ability of automatic tone modeling. In *International Computer Music Conference (ICMC)*, pages 248–255.