IMPROVEMENTS OF CONTINUOUS MODEL FOR MEMORY-BASED AUTOMATIC MUSIC TRANSCRIPTION

Štěpán Albrecht\textsuperscript{1}, Václav Šmídl\textsuperscript{2}

\textsuperscript{1}University of West Bohemia, Plzeň, Czech Republic, albrs@kiv.zcu.cz
\textsuperscript{2} Institute of Information Theory and Automation, Prague, Czech Republic, smidl@utia.cas.cz

ABSTRACT

Automatic music transcription (AMT) is a process recovering the most likely combination of sounds that produced the recorded audio signal. We are concerned with memory-based approach, where the observed signal is modeled as a superposition of sounds from a library. Moreover, we assume that only parts of the sounds can be played. The number of possible combinations is excessive and exact estimation is computationally prohibitive. We propose to transform the original discrete-event model into a less restricted parametrization and impose the constraints in a soft way via prior information. The resulting model is a non-linear state-space model with Gaussian disturbances. The posterior estimates are evaluated by the extended Kalman filter. Performance of the model is studied in simulation and it is shown that it outperforms previously published methods.

1. INTRODUCTION

Automatic music transcription (AMT) is a process of decomposing recorded music signal into a sequence of higher-level sound events. The entire AMT—i.e. resolving pitch, loudness, timing and instrument of all sound events in an input audio music signal [6]—is not theoretically possible in general [6], therefore practical AMT has to be restricted to a specific scenario. Commonly used scenarios are memory-based and data-based AMT. The former utilizes sound models corresponding to a certain musical instrument sound (allowing to identify the instruments), the latter utilizes only rules which hold in general. We are concerned with a special case of memory-based AMT. Kashino’s transcription system [9] is another system that is considered as an entire memory-based AMT system in the sense of [6].

Intuitively, the problem can be understood as an ‘inverse music sequencer’, Fig. 1. Music sequencers have a pre-recorded library of sounds (sound components) which are combined together to create music signal. Input to the sequencer is a MIDI file which contains information about beginning of music events in time, their duration, IDs of sounds (in our case the pre-recorded sound components), their amplitude and modification type. Component modification(s)—e.g. component truncation or pitch shifting—were designed to reduce the size of the pre-recorded library. In this paper we consider only component truncation as a possible modification. Output of the sequencer is the audio signal. Input of our “inverse music sequencer” is the recorded music signal and its output is the estimated (transcribed) MIDI-like representation of music events.

The sequencer composes the output from sounds stored in the library of \( K \) sounds. Each sound is composed of \( L_k \) frames, which are supposed to be played after each other.

![Figure 1: Principle of a music sequencer. The range of active frames \( p_{kk} \) is yellowed. Note that the amplitudes are the same for all events in a track \( k \) (represented by squares of the same color).](image-url)

The input events are defined by: (i) index of the sound to play, \( k = [1, \ldots, K] \), (ii) truncation of the sound, i.e. beginning of the range of frames from the \( k \)th sound to play, \( p \), and (iii) amplitude of the sound, \( 0 \leq g \leq 1 \). We assume that each sound can be played only once at time \( t \). The output sound is then:

\[
y_t = \sum_{k \in \mathcal{K}_t} g_k f(k_k, p_k, t),
\]

where \( y_t \) is the \( \phi \)-dimensional vector of measurements at time \( t \) composed of either time- or frequency-representation of the input music signal segment (frame); \( \kappa \) denotes ID of the event from the set of events active at time \( t \), \( \mathcal{K}_t \subset [1, \ldots, K] \). Function \( f(k_k, p_k, t) \) looks up the frame from range \( p_k \) of the \( k \)th sound that is active in time \( t \), see illustration on Fig. 1.

Model (1) is a suitable representation of a sequencer, however, it is not suitable for the inverse operation since the number of possible configurations of the unknowns \( \mathcal{K}_t \) and \( p_k \) is enormous. Formally, (1) can be written as a sum over all frames

\[
y_t = \sum_{i=1}^{N_L} \alpha_{i,t} g_{i,t} f_{i,t},
\]

where \( N = \sum_{k=1}^{K} L_k \), \( \alpha_{i,t} \in \{0, 1\} \) is equal to 1 if \( i \)th frame, \( f_{i,t} \), is used in (3) and \( g_{i,t} \) is the corresponding amplitude. The values of \( \alpha_{i,t} \) are constrained by the parameters \( \mathcal{K}_t \) and \( p_k \) as follows:
only one frame of the kth sound may be active at time t,
• when the ith frame of the kth sound was active in time
  \( t - 1 \) and t is in \( p_k \), the \((i+1)\)th frame must be active in
time t,
• no frame is active when t is out of \( p_k \).

The number of possible combinations of \( \alpha_{i,t} \) is still enormous,
since we allow arbitrary truncations of the sounds.

Therefore, we propose to relax the hard constraints above and introduce an unconstrained variable \( 0 \leq \alpha_{i,t} \) such that

\[
y_t = \sum_{i=1}^{N} a_{i,t} f_i = F a_t.
\]

where \( F = [f_1, \ldots, f_N] \) and \( a_t = [a_{1,t}, \ldots, a_{N,t}]' \), \( a_{i,t} \) being the amplitude of \( i \)th frame. This relaxation has both advantages and disadvantages.

The advantage is that model (3) is well studied in statistical literature and efficient parameter estimation methods exists for its various variants. For example, linear regression, factor analysis [3], Kalman filtering, matching pursuit [10] and independent component analysis [5] (ICA) arise from (3) by imposing different assumptions on parameters \( a_t \) and \( F \).

These methods are used in music processing, e.g. ICA for blind (unsupervised) source separation (BSS) techniques in monaural input music signals [6]. In this work, we investigate the use of the Kalman filtering approach.

The main disadvantage of the relaxation is that it allows to explain signal \( y_t \) by a combination of frames that are not valid from musical point of view (e.g. it allows to play all frames from one sound at the same time). The original restrictions can be restored in less restrictive form via transition model \( p(a_{t-1}|a_t) \) which needs to be designed. Similar approach was used in [2] where the prior was designed for the whole sequence by optimized combination of priors commonly used in the area. In this paper, we present a new model using only first-order Markov transition model which is obtained by conversion of transition model for discrete events (1) into a form suitable for the continuous model (3).

The paper is organized as follows: model of the signal transition between time frames is presented in the second Section; evaluation of the posterior density implied by the model is presented in the third Section; simulation study that assess performance on the model on music data is in the fourth section.

2. DERIVATION OF THE MODEL

Observation of the signal \( y_t \) are never perfect due to round-off errors and measurement noise. The observation model (3) is used as mean value of Gaussian likelihood function of observations:

\[
p(y_t|a_t, F) = \mathcal{N}(Fa_t, \omega^{-1}I_\phi).
\]

Here, \( \mathcal{N} \) denotes normal distribution of vector argument, \( \omega \) is scalar precision parameter, \( I_\phi \) denotes identity matrix of dimensions \( \phi \times \phi \).

The task is to estimate posterior density of \( a_t \) given available data, \( p(a_t|F, Y_t) \), where \( Y_t = [y_1, \ldots, y_t] \). The constraints on \( \alpha \) will be transformed into Gaussian prior \( p(a_t|a_{t-1}) \), which is parametrized by mean value of size \( N \) and covariance matrix of size \( N \times N \).

2.1 Transformation between \( \alpha_{i,t} \) and \( a_{i,t} \)

We start with a simple transformation between discrete variable \( \alpha_t \) and continuous amplitude \( a_t \), specifically

\[
p(a_{i,t}|\alpha_{i,t}) = \begin{cases} \mathcal{N}(1, \kappa \sigma_1) & \text{if } \alpha_{i,t} = 1, \\ \mathcal{N}(0, \sigma_1) & \text{otherwise.} \end{cases}
\]

Intuitively, zero values of \( \alpha_{i,t} \) (i.e., representation of silence) are mapped on \( a_{i,t} \) which are ‘close to zero’ and \( \alpha_{i,t} = 1 \) (i.e., the loudest sound notation) are mapped to \( a_{i,t} \) close to 1. The closeness is modeled by variance parameter \( \sigma_1 \). Since we allow lower amplitudes of the tone via \( \varphi \), we model variance of the first component of the pdf in (5) to be \( k \) times greater than that of the second component.

Inverse mapping of \( a_{i,t} \) to \( \alpha_{i,t} \) can be obtained by the Bayes rule:

\[
p(\alpha_{i,t}|a_{i,t}) = \frac{p(a_{i,t}|\alpha_{i,t})p(\alpha_{i,t})}{p(a_{i,t})}.
\]

There is no information on prior of \( \alpha_{i,t} \), thus \( p(\alpha_{i,t}) \) is uniform, and for a particular component:

\[
p(\alpha_{i,t} = 0 | a_{i,t}) \propto \frac{1}{2} \sigma_1^{-0.5} \exp(-\frac{-1}{2\sigma_1^2}a_{i,t}^2)
\]

\[
\frac{1}{2} \sigma_1^{-0.5} \left[ \exp(-\frac{-1}{2\sigma_1^2}a_{i,t}^2) + k^{-0.5} \exp(-\frac{-1}{2\sigma_1^2}(1-a_{i,t})^2) \right] = 1 + \frac{1}{1 + k^{-0.5} \exp(-\frac{-1}{2\sigma_1^2}((1-k)a_{i,t}^2 - 2a_{i,t} + 1)})
\]

2.2 Parameter evolution model

In the discrete parametrization (1), the transition between frames can be modeled by a simple Markov transition:

\[
p(\alpha_{i,t-1} | \alpha_{i,t-1-1}) = \begin{cases} \alpha_{i,t-1-1} = 0, & \text{if } \alpha_{i,t-1} = 0, \\ \alpha_{i,t-1-1} = 1, & \text{otherwise.} \end{cases}
\]

where \( \tau_0, \tau_1 \) are constant probabilities that the discrete amplitude is not changed by the transition from \( t - 1 \) to \( t \). This transition model can be combined with (6) as follows:

\[
p(a_{i,t} | a_{i-1,t-1}) = \sum_{\alpha_{i-1,t-1}} p(a_{i-1,t-1} | a_{i-1,t-1})p(\alpha_{i-1,t-1} | a_{i-1,t-1})
\]

However, direct application of this rule would result in prior \( p(a_t) \) being a mixture of \( 4^{NT} \) components which is not computationally tractable. Hence, we project (7) into a single Gaussian density

\[
p(\alpha_{i,t} | a_{i-1,t-1}) = \mathcal{N}(\mu_{i,t-1}, \sigma_{i,t-1})
\]

using geometric merging of probabilities [7], which yields

\[
\sigma_{i,t-1}^{-1} = \alpha_{i,t} (k-1) \tau_0 + \frac{1}{k \sigma_1} + (1 - \alpha_{i,t}) (k-1)(1-\tau_1) + 1
\]

\[
\mu_{i,t-1} = \frac{\alpha_{i,t} (1-\tau_0 + \tau_1) + (1 - \alpha_{i,t}) \tau_1}{k \sigma_1}
\]

\[
\alpha_{i,t} (1-\tau_0 + \tau_1) + 1
\]

\[
\alpha_{i,t} (k-1)(\tau_0 + \tau_1 - 1) + (k-1)(1-\tau_1) + 1
\]
The state-space model derived in Section 2 is strongly non-linear model with Gaussian disturbances. There is a range of techniques for Bayesian filtering, such as particle filters [4], extended Kalman filters, and others. The model was derived using projection into Gaussian densities, hence a filter derived using the standard EKF as follows:

\[
R_y = F^T P_{t-1} F + \sigma_0 I,
\]
\[
K = P_{t-1} F R_y^{-1} \quad (15)
\]
\[
P_{t|t} = P_{t-1} - P_{t-1} F R_{y}^{-1} F^T P_{t-1},
\]
\[
P_t = AP_{t-1}A + Q_t(\hat{a}_{t-1}).
\]

Here, \(A = \frac{d}{d\hat{a}_{t-1}} h(\hat{a}_{t-1})\) which is a sparse matrix composed of derivatives of \(\mu_{ij}\) (10)

\[
\frac{d}{d\hat{a}_{t}} \mu_{i,j} = \frac{(-1 + t_0 + t_1) \sqrt{k^2(\epsilon + (k-2t_0 + t_1))}}{k} \cdot (k^{-1})
\]
\[
\epsilon = \exp \left( -\frac{1}{2} \frac{(k-1) a_{t-1}^2 + 2 a_{t-1} - 1}{k \sigma_1} \right)
\]

Note that \(Q_t(\hat{a}_{t-1})\) in (11) was replaced by \(Q_t(\hat{a}_{t-1})\) in (15). This change is required since EKF does not allow covariance matrices to be function of the state variable. We conjecture that this is an acceptable approximation.

3. BAYESIAN FILTERING OF THE MODEL

The simulated data were generated from piano midi files. Each note was represented by pitch, onset time, duration and offset in the sound library. The offset is a non-standard extension of the midi format. The corresponding amplitude matrix and the observed audio signal were generated using model (1). Midi notes that were not available in the library of sounds were omitted. For testing purposes, 61 library sounds (corresponding to midi notes 36—96) were used, each of them having 10 frames. Each frame contained 4096 samples at 44.1 kHz sample rate, represented by the magnitude spectrum. For training of the nuisance parameters, only 36 (midi notes 45—80) sounds were considered. Thus, there were 610 and 360 frames in the testing and training library, respectively. The sounds assigned to the piano midi events were obtained by a harmonic tone synthesizer [1] which produce tones with sharp attack and inner frames of different loudness, however, the frames were significantly similar to each other. Hence, the audio signal generated by the first
frame at low amplitude is remarkably similar e.g. to that of
the third frame at higher amplitude. This is a challenge
for estimation, since the likelihood model alone can not pro-
perly distinguish those two cases and good model of the prior
is required to obtain good performance.

The proposed model contains nuisance parameters
\( \delta_i = [\sigma_k, k, \tau_0, \nu_1, \nu_2, \nu_3, \nu_4] \) in the apriori part and \( \omega_0 \) in the likelihood.
These were optimized by Matlab function fminsearch using
the following criteria: (i) a measure similar to the total
relative sound-to-distortion ratio \( R \) that read:

\[
SDR = 10 \log_{10} \frac{\sum [b \cdot F_{acoust} - a_i]^2}{\sum [y - b \cdot F_{acoust} a_i]^2}, \tag{16}
\]
where \( b \) is a scalar fitting \( b \cdot A = A_{reference} + noise \) accord-
ing to MMSE, and \( F_{acoust} \) is the matrix of frames in acoustic
form; and (ii) a hit-measure: \( m = hits - 0.5 \cdot (falsepositive +
\) false-negative). Model nuisance parameters were trained on
51 frame long signal of one of Debussy’s preludes and tested on
582 frame long concatenation of short excerpts of Mozart
and Debussy. In the training phase, 51 units were filtered by
the Kalman filter to a selected optimization criteria value,
frame by frame with no overlap. The SDR criteria was found
to be more suitable for optimization since the hit-measure is
too coarse for the fminsearch optimization. Moreover,
the hit-measure depends on the amplitude threshold to dis-
tinguish active from non-active \( a_{ij} \) amplitudes whereas the
SDR does not. All results presented in the paper are based on
nuisance parameters optimized for the SDR criterion. In
the testing phase, 58 seconds of music signal containing
1325 active frames were estimated by the Kalman filter.

For comparison, two previously published methods have
been applied to the same data. The first approach, la-
beled ‘maxent’, is based on model (3) with different prior
[2]. The prior is defined by combination of four phenomena:
(1) sparsity, (2) temporal dependence, (3) dissimilarity of simultane-
ous sounds. Combination of these phenomena was governed by
uinfluence parameters \( \delta_2 = [\lambda, \gamma, c, \nu_1, \nu_2] \) which were optimized using the same
fminsearch procedure. The original \( \delta_2 \) from [2] con-
tained additional parameter \( \omega_0 \) which was found to be re-
dundant. The second compared method, labeled ‘NMF’, is
non-negative matrix factorization of the measurement matrix
\( Y \) [6], where the matrix of basis corresponds to \( F \) that is
known. Even though there are ‘NMF’s with various restric-
tions on amplitude matrix, the considered ‘NMF’ transcription
uses no prior knowledge (i.e., no restrictions) to de-
strate the informativeness of the independent measurements.

Results of transcriptions of all three tested methods are
displayed via piano-roll schematics in Figures 3 (detail of
note E 64) and 4 (initial 25 samples of the testing set). Note
that the above mentioned ambiguity of the likelihood is well
manifested in the results of NMF approach (Fig. 4, bottom-
right) which is based only on the likelihood. Models of prior
information (maxent or the current model) improve the esti-
mation results by sharpening around the most likely path.
However, the ambiguity is affecting these methods as well,
since one missed frame may lead to a postponement of the
<p>| Table 1: Comparison of the presented model with previous methods. |
|-----------------------------|-------------------|-----------------|------------------|-------|</p>
<table>
<thead>
<tr>
<th>active frames</th>
<th>hits</th>
<th>false positive</th>
<th>false negative</th>
<th>SDR [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>current model</td>
<td>1325</td>
<td>1219</td>
<td>293</td>
<td>106</td>
</tr>
<tr>
<td>maxent model</td>
<td>1325</td>
<td>1038</td>
<td>257</td>
<td>557</td>
</tr>
<tr>
<td>NMF</td>
<td>1325</td>
<td>1007</td>
<td>1367</td>
<td>218</td>
</tr>
</tbody>
</table>

This led either to their over-estimation or under-estimation
according to the tested data. Over-estimation of the lengths
causes only minor decrease of the SDR values since ampli-
tudes of the tones at their ends are small.

Using library of those 61 sounds, one time unit processing
ranged from 1.5 to 2 with Kalman filter on Core Duo or
Quad Core CPU. Hundred of iterations of the ‘NMF’ algo-
rithm took about 30 seconds, thus the implementation of the
problem solution by the ‘NMF’ was about 50 times faster
than the problem solution by Kalman filter.

5. DISCUSSION AND CONCLUSION

We have presented a new model with continuous parametrisation
for automatic music transcription. The main motivation of the new prior is on-line transcription of
the signal using only first-order Markov transition model.
The underlying model of discrete events was transformed into
continuous version via Gaussian mixture models. Projection of these mixtures into a single Gaussian density
yields non-linear state-space model with Gaussian distur-
bances. Music transcription is then converted into estimation
of the state variable which is achieved by the extended
Kalman filter with a minor modification. The nuisance
parameters were tuned on a small training set, while the final
comparison was performed on a significantly larger data-set.
The results compare favorably to the previously published
approaches. Note that the transcription is obtained on-line, i.e. each point was estimated using only data available up to
the time of the analysis. It can be expected that extensions
using Bayesian smoothing would further improve on the
quality of the estimates. At present, the Kalman gain calcu-
lation is rather expensive—one step takes about one second
in this case—but there is a lot of space for optimization or
approximate evaluation employing e.g. the ensemble
Kalman filter. Further improvement can be obtained by
extension of the prior to higher-order Markov model. In
this paper, we have considered only transition between the
two consecutive frames in the bank of sounds. Clearly, the approach can be extended for 3 and more frames.

Acknowledgment

Support of grant GA ČR 102/08/P250 is gratefully acknowledged.

REFERENCES