Automatic Music Transcription via Music Component Identification

Štěpán Albrecht
University of West Bohemia, 30614 Plzeň, Czech Republic, Email: albrs@kiv.zcu.cz

Introduction

Automatic music transcription (AMT) [2] is a process of decomposing recorded music signal into a sequence of higher-level sound events. We deal with entire AMT. The entire AMT implies resolving pitch, loudness, timing and instrument of all sound events in an input audio music signal. It is not theoretically possible therefore the practical entire AMT is focused on detection of most of the intrinsic music information. The scenario of this paper follows the inverse operation of the music sequencer (MS). It is a memory-based AMT since it utilizes tone library. The MS operates as follows: we have a library of library sounds. These sounds are or are not modified before they are placed into the tracks. The audio content of the tracks is superposed hence the resulting audio signal is obtained. The inverse operation works so that given the complex audio sound and the library audio sounds the procedure identifies inner sounds with the library sounds and determines their modification parameters.

Observation Model

We propose an observation model of the sequencer suitable for sequential processing which considers discrete-evolving amplitudes and library sound truncation as modification types. The model can be represented by equation

$$Y \approx F \cdot A, \tag{1}$$

where Y is representation of the observed audio signal, F denotes the library sounds and A is the representation of amplitude matrix, in which there is contained all the asked information. In A, the vertical dimension denotes the discrete time while the horizontal dimension represents the frame indices in the sound library. The observation matrix Y is obtained so that the observed audio signal is cut into segments, called frames, which are processed by the magnitude discrete Fourier transform and placed as columns into the matrix Y. The library sounds are concatenated one after another and the audio signal is processed frame-wise identically to the obtaining of Y, this yields the matrix F. Given Y and F in such forms, the simultaneous sounds in the observed music signal produce diagonal sequences in the matrix A, see Fig. 1. Number of possible configurations of desired parameters (amplitudes, truncation parameters) which can be estimated given the model (1) is enormous. The number of configurations can be reduced by application of music principles.

Application of Music Principles

The applied music principles are known from the audio source separation problems [4]. (A) Sparsity – amplitude matrix A must be sparse; (B) temporal dependence – number of interruptions in frame sequences in A must be small, see in Fig. 1. Using Bayesian approach the music principles can be applied so that they are formulated as a priori p.d.f. $p(A)$.

Sparsity (A)

The amplitudes are considered to be close to silence and the loudest sounding. The sparsity p.d.f. $p(a_{i,t}|\alpha_{i,t})$ refers what is the probability that the current frame activity is silence, $\alpha_{i,t} = 0$, or loudest sounding, $\alpha_{i,t} = 1$. Also the probability of silence is assumed to be higher than of loudest sounding due to the k-times multiplication of the variance σ_1 in the formula:

$$p(a_{i,t}|\alpha_{i,t}) \quad = \quad \begin{cases} \mathcal{N}(0, \sigma_1) & , \alpha_{i,t} = 0 \\ \mathcal{N}(1, k \cdot \sigma_1) & , \alpha_{i,t} = 1 \end{cases}$$

Temporal Dependence (B-1), (B-2)

The (B-1) part $p(\alpha_{i,t}|\alpha_{i-1,t-1})$ is the expectation of the silence or loud sound at time t given the silence or loud sound at time $t-1$. This is given by the transition table, where transitions $p(\alpha_{i,t} = 0|\alpha_{i-1,t-1} = 0) = \tau_0$ and $p(\alpha_{i,t} = 1|\alpha_{i-1,t-1} = 1) = \tau_1$. The (B-2) part $p(\alpha_{i-1,t-1}|a_{i-1,t-1})$ implies probability of the silence and loud sound in the time $t-1$ given a true value of the amplitude at time $t-1$. This was calculated from the geometrical merge of (A) and (B-1), in particular we got: $p(\alpha_{i-1,t-1} = 0|a_{i-1,t-1}) = \frac{1-a_{i-1,t-1}}{(k-1)a_{i-1,t-1}+1}$ and $p(\alpha_{i-1,t-1} = 1|a_{i-1,t-1}) = 1 - p(\alpha_{i-1,t-1} = 0|a_{i-1,t-1})$.

Merging (A), (B-1), (B-2)

The music principles are merged [3] together in order to produce one single Gaussian density for one frame amplitude which mean and variance are dependent on the previous frame amplitude.

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

$$= \begin{cases} p(\alpha_{i-1,t-1} = 0|a_{i-1,t-1}) & * \begin{cases} \tau_0 & * \quad \mathcal{N}(0, \sigma_1) \\ (1-\tau_0) & * \quad \mathcal{N}(1, k\sigma_1) \end{cases} \\ p(\alpha_{i-1,t-1} = 1|a_{i-1,t-1}) & * \begin{cases} \tau_1 & * \quad \mathcal{N}(1, k\sigma_1) \\ (1-\tau_1) & * \quad \mathcal{N}(0, \sigma_1) \end{cases} \end{cases}$$

The star operation * denotes "is exponent of" operator, the geometrical merging (multiplication) is represented by curly brackets. The merging itself produces one combined Gaussian from two Gaussians [3].

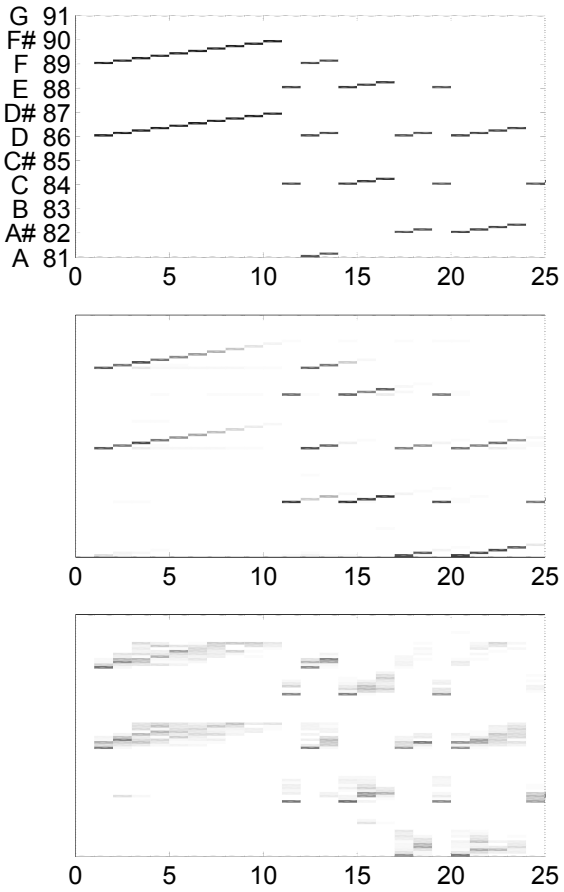

Figure 1: Example of referenced and transcribed piece of polyphonic music. Top: reference-original music excerpt; middle: the excerpt transcribed by the current model with optimized nuissance parameters, bottom: transcription by NMF without any constraints. Vertical axis denotes tone with the due midi keys, the horizontal denotes discrete time (time units).

Transcription Algorithm

The aforementioned model equations can be summarized into

$$p(a_t|a_{t-1}) = \mathcal{N}(\mu(a_{t-1}), \Sigma(\hat{a}_{t-1})), \quad (2)$$

$$p(y_t|a_t) = \mathcal{N}(Fa_t, \omega^{-1} I_\phi). \quad (3)$$

where $\omega^{-1} I_\phi$ is the diagonal covariance matrix of the observation equation and $\Sigma(\hat{a}_{t-1})$ is the diagonal covariance matrix estimate. Such a model is Markovian, Gaussian, non-linear in continuous state space evolving. As an appropriate estimation algorithm for a_t we used the extended Kalman filter (EKF). The EKF iteration can be described as follows:

$$EKF([a_{t-1}, M_{t-1}, \mu(a_{t-1}), \Sigma(\hat{a}_{t-1}), F, \omega \cdot I_\phi]) \longrightarrow [a_t, M_t, \mu(a_t), \Sigma(\hat{a}_t)], \quad (4)$$

where $M_t = \frac{d}{da_{t-1}} \mu(a_{t-1})$.

Experiments

There were 36 synthesized harmonic complex sounds in the sound library, 10 frames in each, one frame length was of 4096 samples at 44.1 kHz, frames did not overlap, values in Y, F, A were between zero and one. This produced covariance matrices of size 360. The frame spectra within one tone were selected to be very similar to each other in order to verify the impact of the a priori part of the model.

As the evaluation measures we proposed the hit measure given as $HM = hits–0.5.(falsepositive + falsenegative)$ and relative sound-to-distortion ratio $SDR = 10log_{10} \frac{\sum_t [b \cdot Fa(t)]^2}{\sum_t [y(t) - b \cdot Fa(t)]^2}$ where b was a scalar minimizing $Y - F \cdot A$ according to the MMSE criterion. Training by Matlab fminsearch was aimed at optimization of 7 nuisance parameters $[\omega, \sigma_1, k, \tau_0, \tau_1, c, q] = [1.21, 10^{-5}, 3.3, 0.99, 0.48, 0.04, 0.003]$, where c, q represents constants in μ, Σ respectively on positions of library sounds beginnings (no temporal a priori information). The training was performed on synthesized [1] MIDI classical music recording of length $T = 51$ frames, containing 200 active (sounding) frames. Around 300 optimization cycles run, better optimization was accomplished with the SDR instead of the hit measure. The transcription samples of the proposed approach and of the non-negative matrix factorization (NMF) are depicted in Fig. 1. The NMF processing can be understand as the music transcription without the applied music principles.

Conclusion and Future Work

Bayesian, Markovian, continuous, gaussian, non-linear model for inverse operation of music sequencer, that is, for memory-based automatic transcription of music was presented. As the transcribing algorithm the extended Kalman filter was used. The approach appeal resides in ability to identify the source and find its modification parameters. Difference between the transcription of synthesized recordings by non-negative matrix factorization and by the transcription utilizing the proposed model was depicted. It can be seen that the music principles such as sparsity and temporal dependence (proposed model) improve transcription over the estimation using the observation model only (NMF). Our future work will be aimed at (1) smoothing techniques in Kalman filtering and also at (2) transition from continuous state-space model to a discrete state-space (resulting in hidden Markov model).

References

[1] http://www.kenschutte.com/midi.

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

[3] Ch. Genest and J. V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1:114–148, 1986.

[4] T Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. In *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, pages 1066–1074, 2007.