Context-free 2D tree structure model of musical notes for Bayesian modeling of polyphonic spectrograms

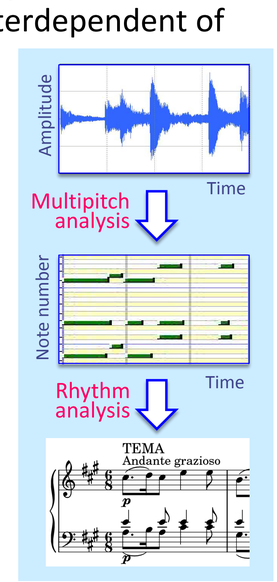
Hirokazu Kameoka[1,2], Kazuki Ochiai[1], Masahiro Nakano[1], Masato Tsuchiya[1], Shigeki Sagayama[1]

[1] The University of Tokyo, [2] NTT Communication Science Laboratories

Abstract: This paper proposes a Bayesian model for automatic music transcription. Automatic music transcription involves several subproblems that are interdependent of each other: multiple fundamental frequency estimation, onset detection, and rhythm / tempo recognition. We formulate a unified generative model, which could be used to jointly solve the problems of determining the pitch and onset time of each musical note, the rhythm and the overall tempo variation of a piece of music, with a Bayesian approach.

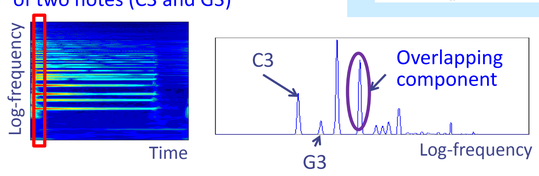
1. Introduction

- Automatic music transcription consists of subproblems that are interdependent of each other:
  - Multipitch analysis
    - Audio source separation
    - Multiple F_0 estimation
    - Onset/offset estimation
  - Rhythm analysis
    - Note value estimation
    - Tempo/beat estimation



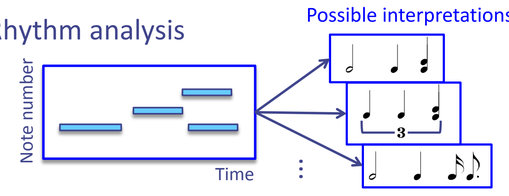
- Ambiguities involved in these subproblems:
  - Multipitch analysis

Spectrogram of mixture of two notes (C3 and G3)



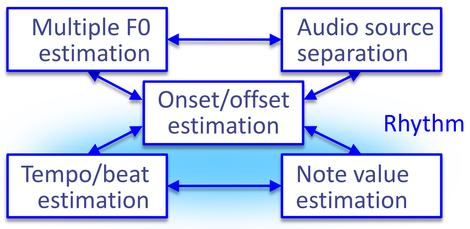
- Since each note consists of many overtones, there can be multiple interpretations of which notes are present in a mixture.

  - Rhythm analysis



- Since we cannot define a note value without having a notion for tempo and vice versa, there can be infinite interpretations regarding what the intended rhythm was and how the tempo varied.

- "Chicken-and-egg" interdependencies



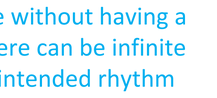
- Simultaneous estimation is generally preferable when several estimation problems are interdependent!

2. Motivation

- How do humans recognize music ?

Humans can recognize music by combining

- hypothesis made according to the observation about when and which notes are active, and



- common sense about how music and music performance should be (e.g., how onset occurrences of notes should be temporally structured, how tempo varies in time)

- We propose to model the generative process of an entire spectrogram of a piece of music by formulating the following three sub-processes and combining them into one process:

①subprocess by which a tempo curve is generated,
②subprocess by which a set of note onset positions (in terms of the relative time) is generated, and
③subprocess by which spectrogram is generated

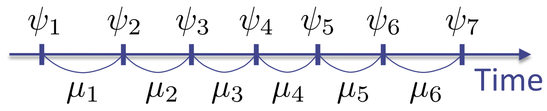
Parameter inference given the observation would then give a musically likely interpretation of what is actually happening in the spectrogram (i.e., a musical score).

3. Generative Model

① Subprocess for generating tempo curve

If we consider 16 divisions per quarter note, the duration of 40 ticks corresponds to two-and-a-half beats.

- Tick d : relative time notion (integer) (the number of discrete divisions a quarter note has been split into)
- Instantaneous (or local) tempo μ_d : the real duration (in units of seconds) corresponding to the interval between *d* and *d* + 1 ticks
- ψ_d : the absolute time (in units of seconds) corresponding to d ticks

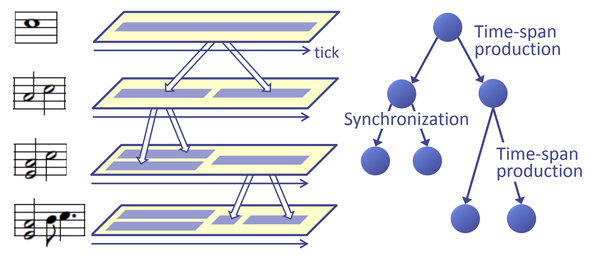


$(\mu_1, \ldots, \mu_D)$ can be regarded as the overall tempo curve of a piece of music

- To ensure smooth overall change in tempo, we assume:

$$\mu_d | \mu_{d-1} \sim \mathcal{N}(\mu_d; \mu_{d-1}, (\sigma^{\mu})^2) \quad (d = 2, \ldots, D)$$

② Subprocess for generating note onset positions



- S_r : Onset position (in units of ticks) of note *r*
- Music has a 2D hierarchical structure, consisting of
  - *Time-spanning* structure
  - *Synchronizing* structure

- We introduce extension of probabilistic context-free grammar (PCFG):

For each node n in the parse tree:

$b_n \sim \mathrm{Bernoulli}(b_n; \phi^{\mathrm{T}})$

[Choose either "EMISSION" or "BINARY-PRODUCTION"]

If b_n = EMISSION

$S_r \sim \delta_{S_r, S_n}, \quad L_r \sim \delta_{L_r, L_n}$ Emit terminal symbol

If b_n = BINARY-PRODUCTION

$\rho_n \sim \mathrm{Bernoulli}(\rho_n; \phi^{\mathrm{N}})$

[Choose either "SYNCHRONIZATION" or "TIME-SPANNING"]

If ρ_n = SYNCHRONIZATION

$S_{n_1} \sim \delta_{S_{n_1}, S_n}, \quad S_{n_2} \sim \delta_{S_{n_2}, S_n}$ Produce 2 copies of note *n*

$L_{n_1} \sim \delta_{L_{n_1}, L_n}, \quad L_{n_2} \sim \delta_{L_{n_2}, L_n}$

If ρ_n = TIME-SPANNING

$S_{n_1} \sim \delta_{S_{n_1}, S_n}, \quad S_{n_2} \sim \delta_{S_{n_2}, S_n + L_{n_1}}$

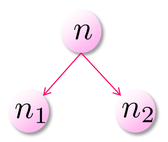
$L_{n_1} \sim \delta_{L_{n_1}, L_n - L_{n_2}}$

$L_{n_2} \sim \mathrm{Discrete}(L_{n_2}; \phi^{\mathrm{B}}_{L_n})$ Split note *n* into two consecutive notes n_1 and n_2

Time-spanning binary production rules

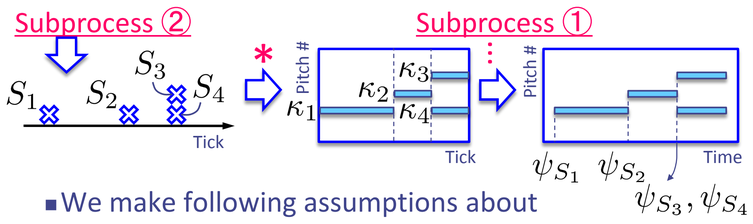
Parent		Child 1	Child 2
𝅝	→	𝅗𝅥	𝅗𝅥
𝅗𝅥	→	𝅗𝅥	𝅘𝅥𝅮
𝅗𝅥	→	𝅘𝅥𝅮	𝅗𝅥
𝅗𝅥	→	𝅘𝅥𝅮	𝅘𝅥𝅮

Parent and children nodes



③ Subprocess for generating music spectrogram $Y_{\omega,t}$ (ω : frequency, t : time)

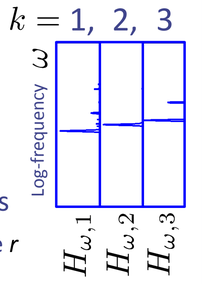
- For each note *r*, we assign pitch index κ_r according to $\kappa_r \sim \mathrm{Discrete}(\kappa_r; \phi^{\mathrm{K}}_r)$ ⋯ ＊



- We make following assumptions about source spectra:

[A1] Each musical note has static spectral profile characterized by particular pitch.
[A2] Magnitude spectrum is represented by superposition of spectra of multiple notes.
[A3] Power of each musical note varies continuously in time.

Spectrum templates with pitch index $k = 1, 2, 3$

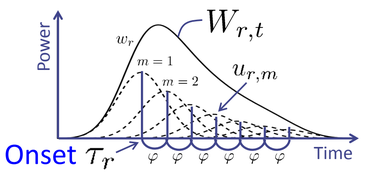


[A1][A2]→ Spectrogram can be described as

$$X_{\omega,t} = \sum_{r=1}^{R} H_{\omega,\kappa_r} W_{r,t}$$

($W_{r,t}$: Power of note *r* at time *t*; H_{ω,κ_r}: Spectrum of note *r*)

[A3]→ We describe power envelope $W_{r,t}$ using a parametric function:



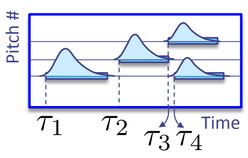
$$W_{r,t} = \sum_{m=1}^{M} G_{r,m,t},$$

$$G_{r,m,t} = \frac{w_r u_{r,m}}{\sqrt{2\pi\varphi}} e^{-(t-(m-1)\varphi-\tau_r)^2/2\varphi^2}$$

- Onset position τ_r of note *r* should be placed near the absolute time into which S_r is converted. To ensure this, we assume $\tau_r \sim \mathcal{N}(\tau_r; \psi_{S_r}, (\sigma^{\tau})^2)$.

- Finally, we assume that observed spectrogram is generated according to:

$$Y_{\omega,t} \sim \mathrm{Poisson}(Y_{\omega,t}; X_{\omega,t})$$



Note: The maximization of the Poisson likelihood with respect to $X_{\omega,t}$ amounts to optimally fitting $X_{\omega,t}$ to $Y_{\omega,t}$ by using the I-divergence as the fitting criterion.

4. Posterior Inference

- Variational inference
  - Goal is to compute $p(\Theta|Y)$

Θ

$w = \{w_r\}_r$	: total energy of note r
$u = \{u_{r,m}\}_{r,m}$	: shape of power envelope of note r
$\tau = \{\tau_r\}_r$	: onset time (sec) of note r
$\kappa = \{\kappa_r\}_r$	: pitch index assigned to note r
$\psi = \{\psi_d\}_d$	: absolute time corresponding to d ticks
$\mu = \{\mu_d\}_d$	: local tempo between d and d+1 ticks
$S = \{S_r\}_r$	: onset position of note r (in ticks)
$L = \{L_r\}_r$	: duration of note r (in ticks)

To obtain exact posterior, we must derive marginal distribution $p(Y)$, but this involves intractable integrals.

- Instead, we consider finding

$$\hat{q}(\Theta) = \underset{q \in \mathcal{Q}}{\mathrm{argmin}}\, \mathrm{KL}(q(\Theta) \| p(\Theta|Y))$$

Kullback-Leibler divergence between $q(\Theta)$ and $p(\Theta|Y)$

Tractable choices of $\mathcal{Q}$ allows us to obtain principled approximate solutions

- Mean-field approximation

$$q(\Theta) = q(C)q(w)q(u)q(\tau, \psi, \mu)q(S, L)$$

- Coordinate ascent
  - By iteratively optimizing one factor q(.) in the mean-field approximation of the posterior at a time while fixing all the other factors, we can find a local optimum.

Probability distributions

$x \sim \delta_{x,y}$ means $x = y$ (with probability 1)

$\mathcal{N}(x; \mu, \sigma) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$\mathrm{Bernoulli}(x; y) = y^x (1-y)^{1-x}$

$\mathrm{Beta}(y; \boldsymbol{z}) \propto y^{z_1 - 1}(1-y)^{z_2 - 1}$

where $x \in \{0, 1\}, 0 \le y \le 1$ and $\boldsymbol{z} = (z_1, z_2)$

$\mathrm{Discrete}(x; \boldsymbol{y}) = y_x \qquad \mathrm{Dirichlet}(\boldsymbol{y}; \boldsymbol{z}) \propto \prod_i y_i^{z_i - 1}$

where $\boldsymbol{y} = (y_1, \ldots, y_I)$ with $y_1 + \cdots + y_I = 1$ and $\boldsymbol{z} = (z_1, \ldots, z_I)$

$\mathrm{Poisson}(y; x) = x^y e^{-x}/y!$

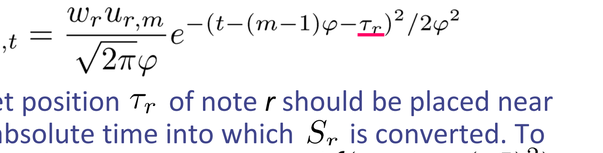
5. Music Transcription Experiments

- Experiment 1
  - Test data: piano recording (RWC-MDB-C-2001 No. 26) Morzart: Piano Sonata No. 11 in A major, K. 331/300i
  - Monaural sampling rate=16kHz
  - Constant-Q transform was used to obtain the observed spectrogram.
  - Note onsets $\tau_1, \ldots, \tau_R$ were given manually.
  - Result: (b) and (c)

(a) Correct score

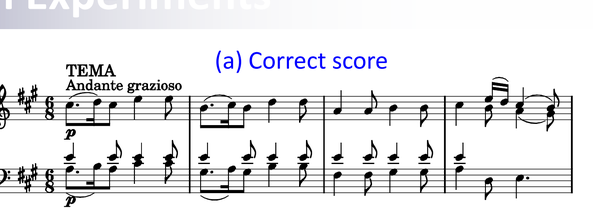
(b) Detected beat locations along with the estimate of $W_{r,t}$

(c) Score transcribed with the proposed method

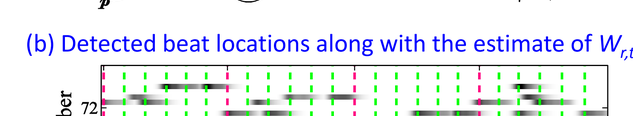


- Experiment 2
  - Next, we applied our method to the same data without providing any information about τ.
  - Result: (d) and (e)

(d) Detected beat locations along with the estimate of $W_{r,t}$



(e) Score transcribed with the proposed method



The result showed that many octave errors had occurred. This kind of error often occurs when there is a mismatch between a spectrum model and an actual spectrum. The validity of the assumptions we have made about the spectra of musical sounds must be carefully examined in the future.

References

[1] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Proc. ICASSP*, Vol. 1, pp. 65-68, 2007.
[2] K. Ochiai, H. Kameoka, and S. Sagayama, "Explicit beat structure modeling for non-negative matrix factorization-based multipitch analysis," in *Proc. ICASSP*, pp. 133-136, 2012.
[3] A. T. Cemgil, "Bayesian inference in non-negative matrix factorisation models," Technical Report CUED/F-INFENG/TR.609, University of Cambridge, 2008.
[4] M. D. Hoffman, D. M. Blei, and P. R. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. ICML*, pp. 439-446, 2010.
[5] K. Yoshii, and M. Goto, "A nonparametric Bayesian multipitch analyzer based on infinite latent harmonic allocation," *IEEE Trans. Audio, Speech, Language Process.*, Vol. 20, No. 3, pp. 717-730, 2012.
[6] M. Nakano, Y. Ohishi, H. Kameoka, R. Mukai, and K. Kashino, "Bayesian nonparametric music parser," in *Proc. ICASSP*, pp. 461-464, 2012.
[7] P. Liang, S. Petrov, M. I. Jordan, and D. Klein, "The infinite PCFG using hierarchical Dirichlet processes," in *Proc. EMNLP*, pp. 688-697, 2007.
[8] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. on Audio, Speech, Language Process.*, Vol. 15, No. 3, pp. 982-994, 2007.