# HARMONIC-TEMPORAL-STRUCTURED CLUSTERING VIA DETERMINISTIC ANNEALING EM ALGORITHM FOR AUDIO FEATURE EXTRACTION

**Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama**
Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
e-mail: {kameoka,nishi,sagayama}@hil.t.u-tokyo.ac.jp

## ABSTRACT

This paper proposes "harmonic-temporal structured clustering (HTC) method", that allows simultaneous estimation of pitch, intensity, onset, duration, etc., of each underlying source in multi-stream audio signal, which we expect to be an effective feature extraction for MIR systems. STC decomposes the energy patterns diffused in time-frequency space, i.e., a time series of power spectrum, into distinct clusters such that each of them is originated from a single sound stream. It becomes clear that the problem is equivalent to geometrically approximating the observed time series of power spectrum by superimposed harmonic-temporal structured models (HTMs), whose parameters are directly associated with the specific acoustic characteristics. The update equations in DA(Deterministic Annealing)EM algorithm for the optimal parameter convergence are derived by formulating the model with Gaussian kernel representation. The experiment showed promising results, and verified the potential of the proposed method.

**Keywords:** audio feature extraction, multi-pitch estimation, harmonic-temporal structured clustering.

## 1 INTRODUCTION

Automatic audio feature extraction of music signals has been taken as one of the most important topics in recent music processing area, towards developing music information retrieval (MIR) systems. This paper describes a new approach of extracting audio features, e.g., pitch, onset, duration, intensity, timbre and so forth of underlying note events, simultaneously from input multi-stream music signal, based upon bottom-up deterministic model parameter optimization methodology.

Developing reliable multi-pitch analysis algorithm for accurately obtaining these features is of primary importance. Contrary to this requirement, the standard level of the numerous conventional methods for multi-pitch analysis has been considered to be far from a practical use. However, the recent pioneering ideas, e.g., graphical model based (Kashino et al., 1995), filter-bank based (Klapuri et al., 2000), Kalman filtering based (Nishi et al., 1996; Abe and Ando, 2000), multi-agent based (Nakatani, 2002) and parametric signal and spectrum modelings based approaches (Feder and Weinstein, 1988; Chazan et al., 1993; Godsill and Davy, 2002; Goto, 2004; Kameoka et al., 2005) brought remarkable progress. While multi-pitch analysis is, in general, a typical ill-posed problem of extracting necessary information lying beneath an ambiguous observation, most of these methods made the problem solvable basically by dealing with frequency and time dimensions separately: first extract instantaneous pitch likelihoods of concurrent sources at each short-time segment and then interpolate/extrapolate them to build up most likely overall continuous temporal pitch structures of multiple audio streams. In auditory scene analysis (ASA), these two processes in human audition are generally called 'segregation' and 'integration', respectively.

In contrast to the common strategy based on sequential integration of instantaneous pitch likelihoods extracted via segregation process, whose performance depends critically on how precisely the segregation process works, this paper aims to offer yet another framework based on simultaneous estimation of geometric structures in both frequency and time directions of power spectra of underlying sound sources.



## 2 GENERAL FORMULATION

Consider time series of observed power spectrum $W(x,t)$, where $x$ and $t$ are log-frequency and time, whose domain of definition is

$$D = \{x, t \in \mathbb{R} \mid \Omega_0 \leq x \leq \Omega_1, T_0 \leq t \leq T_1\}. \quad (1)$$

The problem we are dealing with is to decompose this observed pattern into $K$ number of sequential spectral streams, i.e., clusters, such that each is originated from a single distinct source activation. This problem is, obviously, an unsupervised categorization of the energy density $W(x,t)$ at each coordinate $(x,t)$, and is hardly a straightforward issue.

The observed energy density $W(x,t)$ at each coordinate $(x,t)$ is not always completely originated from a single source but rather superposed by energy patterns generated from different sources that are located close to each others in $xt$ plane, making it totally ambiguous. Thus, we shall assume that each energy density $W(x,t)$ has fuzzy membership, i.e., the membership degree $m(k;x,t)$, in k-th cluster. Approximately assuming that observed power spectral densities are the sum of actual power densities of underlying sources, which is not precisely true but acceptable in expectation sense (where phase differences take uniformly random values), $m(k;x,t)$ satisfies

$$\sum_{\forall k} m(k;x,t) = 1, \;\; \forall k, 0 \le m(k;x,t) \le 1. \quad (2)$$

Therefore, $m(k;x,t)W(x,t)$ denotes the decomposed spectral density of the k-th source, i.e., k-th cluster. Let us define $q_k(x,t;\Theta)$ as a function modeling latent distinct spectral stream density of k-th active source (in music, corresponds to a single note event), governed by parameter vector $\Theta$, where the class of the sources mentioned here, in general, includes not only harmonic signals but even white or pink noises or any others, as far as those properties can be well modeled in $q_k(x,t;\Theta)$ with a mathematical representation. Now the function $q_k(x,t;\Theta)$ is what we are to estimate and 'goodness' of the partitioned cluster $m(k;x,t)W(x,t)$ can be measured by a quasi-distance of $m(k;x,t)W(x,t)$ and $q_k(x,t;\Theta)$:

$$\iint_D \underbrace{m(k;x,t)W(x,t)}_{\text{density of cluster } k} \log \frac{m(k;x,t)W(x,t)}{q_k(x,t;\Theta)} dxdt \quad (3)$$

Though defining some other forms for the quasi-distance is certainly possible such like $L^2$ norm, the intention of giving this specific form shown above will become clear in the following descriptions. It is obvious that as $q_k(x,t;\Theta)$ and $m(k;x,t)W(x,t)$ become closer, Eq. 3 approaches zero. Hence a global cost function of the clustering to minimize w.r.t. $\Theta$ is given as

$$J = \sum_{\forall k} \iint_D m(k;x,t)W(x,t) \log \frac{m(k;x,t)W(x,t)}{q_k(x,t;\Theta)} dxdt \quad (4)$$

subjected to

$$\iint_D W(x,t)dxdt = \sum_{\forall k} \iint_D q_k(x,t;\Theta)dxdt = W \quad (5)$$

(to let J be non-negative, cf., Jensen's inequality) where it can be further rewritten as

$$J = -I(\Theta) - \lambda \left( \sum_{\forall k} m(k;x,t) - 1 \right) + \sum_{\forall k} \iint_D m(k;x,t)W(x,t) \log m(k;x,t)W(x,t) dxdt$$

$$I(\Theta) \equiv \sum_{\forall k} \iint_D m(k;x,t)W(x,t) \log q_k(x,t;\Theta) dxdt \quad (6)$$

where $\lambda$ is a Lagrange multiplier. Although minimizing J w.r.t. both $\Theta$ and $m(k;x,t)$ rarely has an analytic solution, it can be monotonically decreased by alternately optimizing $\Theta$ and $m(k;x,t)$, similar to the basic iterative clustering algorithm such as the k-means algorithm. Partial derivative of the integrand in J w.r.t. $m(k;x,t)$ is

$$W(x,t)\left(1 + \log \frac{m(k;x,t)W(x,t)}{q_k(x,t;\Theta)}\right) - \lambda \quad (7)$$

such that setting it zero gives

$$m(k;x,t) = \frac{q_k(x,t;\Theta)}{W(x,t)} \exp\left(\frac{\lambda}{W(x,t)} - 1\right). \quad (8)$$

From Eqs. 2 and 8, we get

$$\lambda = W(x,t)\left(1 - \log \frac{\sum_{\forall k} q_k(x,t;\Theta)}{W(x,t)}\right) \quad (9)$$

such that substituting Eq. 9 in Eq. 8, we finally have the optimal membership degree under fixed $\Theta$, given as

$$\widehat{m}(k;x,t) = \frac{q_k(x,t;\Theta)}{\sum_{\forall k} q_k(x,t;\Theta)}. \quad (10)$$

Substituting Eq. 10 in 4, it becomes exactly the same form as the KL(Kullback-Leibler) divergence between $W(x,t)$ and the sum of $q_k(x,t;\Theta)$ for all k, i.e.,

$$J_{m_k = \widehat{m}_k} = \iint_D W(x,t) \log \frac{W(x,t)}{\sum_{\forall k} q_k(x,t;\Theta)} dxdt \quad (11)$$

such that this clustering can also be understood as a model-based geometric optimal approximation. Another interesting interpretation of this result is that by regarding k as missing data and replacing $q_k(x,t;\Theta)$ with complete data pdf $p(k,x,t|\Theta)$, it proves the convergence of EM(Expectation-Maximization) algorithm from another viewpoint without applying any probability laws. The correspondence to the EM algorithm becomes much clearer by comparing Eqs. 6 and 10 with Q function, given by

$$Q(\Theta, \widetilde{\Theta}) = \sum_{\forall k} \iint_D \overbrace{p(k|x,t,\Theta)}^{\text{missing data pdf}} \overbrace{W(x,t)}^{\text{observed pdf}} \log \overbrace{p(k,x,t|\widetilde{\Theta})}^{\text{complete data pdf}} dxdt$$

$$p(k|x,t,\Theta) = \frac{p(k,x,t|\Theta)}{p(x,t|\Theta)} = \frac{p(k,x,t|\Theta)}{\sum_{\forall k} p(k,x,t|\Theta)}$$

where $k, x, t \in \Omega$ (probabilistic variable) and

$$\iint_D W(x,t)dxdt = 1, \;\; \sum_{\forall k} \iint_D p(k,x,t|\Theta)dxdt = 1.$$

Under fixed membership degree $m(k;x,t)$, on the other hand, parameter $\Theta$ can be updated by

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmin}}\, J = \underset{\Theta}{\operatorname{argmax}}\, I(\Theta) \left( = \underset{\widetilde{\Theta}}{\operatorname{argmax}}\, Q(\Theta, \widetilde{\Theta}) \right) \quad (12)$$

according to the specific form of $q_k(x,t;\Theta)$, which will be formulated in the next section. We call the methodology for multi-pitch analysis based on this general formulation 'spectro-temporal structured clustering (STC)'.

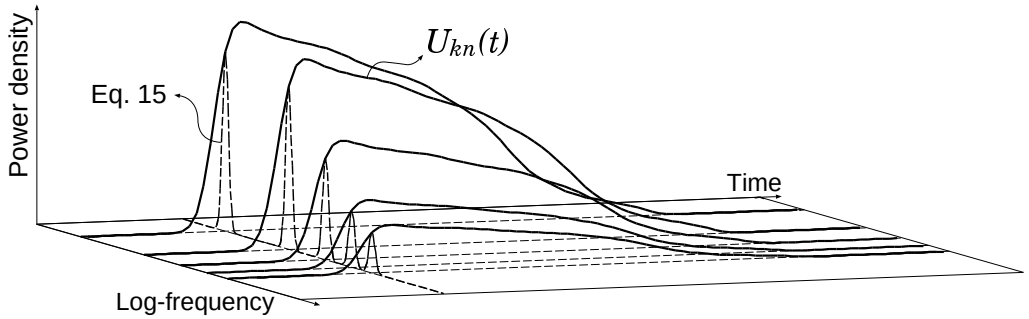



Figure 1: k-th harmonic-temporal structured model (HTM) $q_k(x, t; \Theta)$ (Eq. 17)

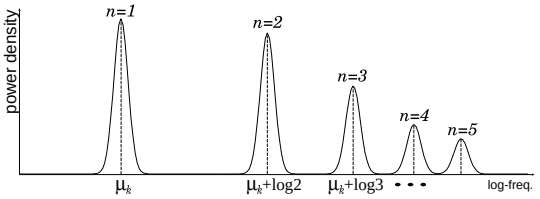



Figure 2: Cutting plane of $q_k(x, t; \Theta)$ at time t (Eq. 15)

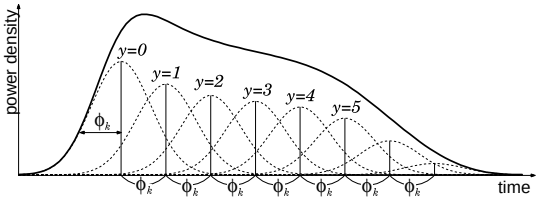



Figure 3: Power envelope function $U_{kn}(t)$ (Eq. 19)

# 3 HTC FORMULATION

## 3.1 MODEL REPRESENTATION

In this section, a mathematical form of $q_k(x, t; \Theta)$ is described. Now let us focus only on harmonic signal, that has pitch or fundamental frequency ($F_0$), through the rest of this paper, given that music audio feature extraction is indeed what we are practically aiming for. Let us call the model particularly limited to harmonic signals 'harmonic-temporal structured model (HTM)'. Suppose the fundamental log-frequency trajectory during a single source activation is expressed with a polynomial

$$\mu_k(t) = \mu_{k0} + \mu_{k1} t + \mu_{k2} t^2 + \cdots \tag{13}$$

(imagine vibrato or glissando), a cutting plane of $q_k(x, t; \Theta)$ at particular time t shall form a pure harmonic structure of fundamental log-frequency $\mu_k(t)$ (see Fig. 2).

Frequency and power of each partial in harmonic structure yield continuous curves along time. Given fundamental log-frequency trajectory $\mu_k(t)$ in k-th HTM, frequency trajectory of the n-th partial is $\mu_k(t) + \log n$. Now if each partial distribution is approximated by a Gaussian function, which is a quite convincing modeling especially when spectra are obtained by Gabor wavelet transform, and suppose power envelope curve of n-th partial is denoted by $U_{kn}(t)$ (presumed to be a function that is normalizable since $q_k(x, t; \Theta)$ has to satisfy Eq. 5),

$$\forall k, \forall n, \int_{-\infty}^{\infty} U_{kn}(t) dt = 1, \tag{14}$$

the power density of the n-th partial in k-th HTM is expressed as a multiplication:

$$U_{kn}(t) \times \underbrace{\frac{\nu_{kn}}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k(t)-\log n)^2}{2{\sigma_k}^2}}}_{\text{weighted Gaussian centered at } x=\mu_k(t)+\log n} \quad (n=1,\cdots,N) \tag{15}$$

where $\sigma_k$ denotes the width of every partial distribution and $\nu_{kn}$ is the relative power of n-th partial, that satisfies

$$\forall k, \sum_{\forall n} \nu_{kn} = 1. \tag{16}$$

Therefore, the power density of k-th HTM, i.e., $q_k(x, t; \Theta)$, as a whole (see Fig. 1) becomes

$$q_k(x, t; \Theta) = w_k \sum_{\forall n} \frac{\nu_{kn} U_{kn}(t)}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k(t)-\log n)^2}{2{\sigma_k}^2}} \tag{17}$$

where $w_k$ indicates the intensity of the k-th source. Further, superposition of K number of spectral streams, i.e., overall density of the model for given observation $W(x, t)$, shall be expressed as a sum of HTMs,

$$L(x, t; \Theta) = \sum_{\forall k} q_k(x, t; \Theta) \quad (k=1,\cdots,K) \tag{18}$$

Since developing general algorithm for music audio feature extraction that appropriately works even if any instruments are used is a completely 'blind' problem, it is perhaps wise not to limit the class of the power envelope function $U_{kn}(t)$ to a model valid only for a particular physical sound production mechanism. Thus, the general modeling of $U_{kn}(t)$ is one of the core parts of this work.

$U_{kn}(t)$ is supposed to be continuous, non-negative, converging to zero at both ends of the time axis, adaptable to any observed curves and elastic in time direction. Furthermore, to accomplish Eq. 12 and to satisfy Eq. 14, it should be differentiable and infinite integrable. Finding non-linear function satisfying these requirements at the same time is hardly simple, however, we came up to formulating it with an original Gaussian kernel function, which is given as

$$U_{kn}(t) = \sum_{y=0}^{Y-1} \frac{u_{kny}}{\sqrt{2\pi}\phi_{kn}} \exp\left(-\frac{(t-\tau_k-y\phi_{kn})^2}{2{\phi_{kn}}^2}\right) \tag{19}$$

where $\tau_k$ is the center of the forefront Gaussian, that should be treated as an onset time estimate, $u_{kny}$ is the

Table 1: The feature parameters that can possibly be useful for MIR systems

| denotation | physical meanings |
|---|---|
| $\mu_k(t)$ | pitch trajectory during k-th source activation (0-order polynomial would be a reasonable way to use) |
| $w_k$ | intensity of k-th active source |
| $\nu_{kn}$ | relative energy of n-th partial stream (perhaps useful as a timbre feature) |
| $u_{kny}$ | decisive element characterizing the shape of power envelope curve of n-th partial stream |
| $\tau_k$ | onset time of k-th source activation |
| $Y\phi_k$ | duration of k-th source activation |

weight parameter multiplied to each kernel, allowing the function to be adaptable to various shapes (when $\phi_{kn} \to 0$ and $Y \to \infty$, this function becomes principally transformable to fit any non-negative functions), that satisfies

$$\forall k, \forall n, \sum_{\forall y} u_{kny} = 1 \tag{20}$$

The remarkable originality in this function is that the $Y$ number of Gaussian kernels are centered with the equal interval of their common standard deviation parameter $\phi_{kn}$ (see Fig. 3). It may be quite unfamiliar to find the standard deviation parameter $\phi_{kn}$ also within the numerator inside the exponential of Gaussian. This specific feature makes $U_{kn}(t)$ a linear elastic function allowing various durations of note events and never lets each kernel be isolated, so that $U_{kn}(t)$ is always ensured to be a smooth envelope.

The parameters of the HTM to estimate, that can be essentially useful as acoustic features available for MIR systems, are listed in Table 3.1. One may realize that most of the parameters in HTM directly reflect decisive features characterizing music performances.

### 3.2 KERNEL SUBCLUSTERING

As the HTM is specified as a kernel function representation, solving Eq. 12 can be mathematically simplified by further breaking down each cluster into $\{n, y\}$-labeled subclusters, associated with the kernel functions.

$q_k(x, t; \Theta)$ can be simply broken down into a sum of $\{k, n, y\}$-labeled kernel density $S_{kny}(x, t; \Theta)$,

$$\begin{aligned} q_k(x,t;\Theta) &= w_k \sum_{\forall n} \left\{ \underbrace{\frac{\nu_{kn}}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k(t)-\log n)^2}{2\sigma_k^2}}}_{=H_{kn}(x,t)} \right. \\ &\qquad \left. \left( \sum_{\forall y} \underbrace{\frac{u_{kny}}{\sqrt{2\pi}\phi_{kn}} e^{-\frac{(t-\tau_k-y\phi_{kn})^2}{2\phi_{kn}^2}}}_{=E_{kny}(t)} \right) \right\} \\ &= \sum_{\forall n} \sum_{\forall y} \underbrace{w_k H_{kn}(x,t) E_{kny}(t)}_{=S_{kny}(x,t;\Theta)} \end{aligned} \tag{21}$$

Introducing $m(n, y; k, x, t)$, membership degree of the k-th partitioned cluster $m(k; x, t)W(x, t)$ in the $\{n, y\}$-labeled subcluster, that satisfies

$$\forall k, \sum_{\forall n} \sum_{\forall y} m(n, y; k, x, t) = 1, \quad 0 \le m(n, y; k, x, t) \le 1,$$

we have the inequality

$$\begin{aligned} J_k &\equiv \iint_D m(k; x, t) W(x, t) \log \frac{m(k; x, t) W(x, t)}{\sum_{\forall n, \forall y} S_{kny}(x, t; \Theta)} dx dt \\ &\le \widetilde{J}_k \equiv \sum_{\forall n, \forall y} \iint_D m(k; x, t) m(n, y; k, x, t) W(x, t) \\ &\qquad \log \frac{m(k; x, t) m(n, y; k, x, t) W(x, t)}{S_{kny}(x, t; \Theta)} dx dt \end{aligned} \tag{22}$$

where the equality holds when

$$m(n, y; k, x, t) = \frac{S_{kny}(x, t; \Theta)}{\sum_{\forall n} \sum_{\forall y} S_{kny}(x, t; \Theta)} \tag{23}$$

(the proof will be omitted since it can be easily proved by following the same derivation as in section 2). Using Eq. 23 as a subcluster membership degree, one can make $\widetilde{J} = \sum_{\forall k} \widetilde{J}_k$ equivalent to the global cost function $J = \sum_{\forall k} J_k$ and minimizing $\widetilde{J}$ obviously offers absolutely better prospect for yielding the solution to Eq. 12 than directly solving Eq. 12. Accordingly, the parameter update equation shall be derived by

$$\begin{aligned} \widehat{\Theta} &= \operatorname*{argmin}_{\Theta} J \\ &\Leftrightarrow \operatorname*{argmax}_{\Theta} \sum_{\forall k} \sum_{\forall n, \forall y} \iint_D \overbrace{m(k; x, t) m(n, y; k, x, t)}^{\equiv m(k, n, y; x, t)} W(x, t) \\ &\qquad \log S_{kny}(x, t; \Theta) \; dx dt \end{aligned} \tag{24}$$

## 4 INTERPRETATION AS MAP

### 4.1 PRIOR DISTRIBUTION ASSUMPTION

Suppose $W(x, t)$ is an observed pdf and $L(x, t; \Theta) = \sum_{\forall k} q_k(x, t; \Theta)$ is a parameter conditional pdf (i.e., model likelihood density) in Eq. 11, one can also interpret our ultimate objective as being equivalent to maximizing expectation of the log-likelihood (what is generally called as maximum likelihood problem), i.e.,

$$\begin{aligned} \widehat{\Theta}_{ML} &= \operatorname*{argmin}_{\Theta} J_{m_k = \widehat{w}_k} \\ &\Leftrightarrow \operatorname*{argmax}_{\Theta} \left\langle \log L(x, t; \Theta) \right\rangle_{W(x,t)} \end{aligned} \tag{25}$$

w.r.t. $\Theta$, where $\langle \cdot \rangle_\Omega$ refers to as an expectation. Interpreting in this way, it is natural to expand our problem to MAP(*Maximum A Posteriori*) estimation by introducing

prior distributions $p(\Theta)$ on the parameters, so that from the Bayes theorem, optimal parameters under prior constraints to estimate could be found by maximizing the expectation of the logarithmic posterior probability, given by

$$\widehat{\Theta}_{\mathrm{MAP}} = \underset{\Theta}{\operatorname{argmax}} \left\langle \log L(x,t;\Theta) + \log p(\Theta) \right\rangle_{W(x,t)} \quad (26)$$

Prior distribution assumption often plays a big role in contributing to prevent model estimates from overfitting. For example, we do not of course wish for a model of octave errored pitch estimate that corresponds to subharmonics of the true pitch, where in this kind of situation, the model tends to be estimated as an 'abnormal' timbre (such that all partial components except particular ones are zero). This can be, however, avoided by assuming prior distribution on $v_{kn}$ so as to prevent the model from 'abnormal' timbre estimates. For another example, we do not indeed want several models to build a power envelope curve that is supposed to be originated from a single source activation (otherwise it would be estimated as several onset times). This situation could also be reduced by assuming prior distribution on $u_{kny}$.

We apply the prior distribution proposed by Goto (2004), which is explicitly given by:

$$\begin{cases} p(\boldsymbol{v}_k) & \equiv \dfrac{1}{Z_v} \exp\left( - d_v \sum\limits_{\forall n} \bar{v}_n \log \dfrac{\bar{v}_n}{v_{kn}} \right) \\ p(\boldsymbol{u}_{kn}) & \equiv \dfrac{1}{Z_u} \exp\left( - d_u \sum\limits_{\forall y} \bar{u}_y \log \dfrac{\bar{u}_y}{u_{kny}} \right) \end{cases} \quad (27)$$

$$\sum_{\forall n} \bar{v}_n = 1, \quad \sum_{\forall y} \bar{u}_y = 1 \quad (28)$$

where $\bar{r}_n$ and $\bar{c}_y$ are the most preferred 'expected' values of $v_{kn}$ and $u_{kny}$, $d_r$ and $d_c$ are contribution degrees of the priors and $Z_r$ and $Z_c$ are normalization factors. Both $p(\boldsymbol{r}_k)$ and $p(\boldsymbol{c}_{kn})$ take maximum value when $v_{kn} = \bar{r}_n$ and $u_{kny} = \bar{c}_y$ for all $n$ and $y$. When $d_r$ and $d_c$ are zero, $p(\boldsymbol{r}_k)$ and $p(\boldsymbol{c}_{kn})$ become uniform (noninformative-prior) distributions. The advantage of using this particular form is in a considerable simplification of calculating Lagrange multipliers in maximizing Eq. 24 without affecting its substance. Note that Dirichlet distribution is also practically applicable.

Given that $\gamma_w$, $\gamma_r^k$ and $\gamma_c^{kn}$ are Lagrange multipliers for $w_k$, $v_{kn}$ and $u_{kny}$, thus what we are to solve to derive the update equation of $\Theta$ under prior constraint is

$$\begin{aligned} \widehat{\Theta}_{\mathrm{MAP}} = \underset{\Theta}{\operatorname{argmax}} \sum_{\forall k} \Bigg( \Bigg( \sum_{\forall n} \sum_{\forall y} \iint_D m(k,n,y;x,t) \\ W(x,t) \log S_{kny}(x,t;\Theta) dx dt \Bigg) \\ - d_v \sum_{\forall n} \bar{v}_n \log \frac{\bar{v}_n}{v_{kn}} - d_u \sum_{\forall n} \sum_{\forall y} \bar{u}_y \log \frac{\bar{u}_y}{u_{kny}} \\ - \gamma_v^{(k)} \Bigg( \sum_{\forall n} v_{kn} - 1 \Bigg) - \sum_{\forall n} \gamma_u^{(kn)} \Bigg( \sum_{\forall y} u_{kny} - 1 \Bigg) \Bigg) \\ - \gamma_w \Bigg( \sum_{\forall k} w_k - 1 \Bigg) \end{aligned} \quad (29)$$

### 4.2 DAEM ALGORITHM (Ueda and Nakano, 1998)

One of the crucial problems in any traditional iterative parameter estmation algorithms is that, the more models become complex, the more likely they cause the estimates to be trapped into local minima/maxima. There have been many efforts for such difficulty over decades in wide research area. For instance, Deterministic Annealing EM (DAEM) algorithm proposed by Ueda and Nakano (1998) is known to be one of the effective approaches offering stable convergence to global maximum/minimum.

So far we have shown that the iterative procedure of updating $m(k,n,y;x,t) = m(k;x,t)m(n,y;k,x,t)$ and $\Theta$ guarantees the convergence of $\Theta$ to a stationary point. From Eqs. 10 and 23, the subcluster membership $m(k,n,y;x,t)$ should be updated to

$$\begin{aligned} \widehat{m}(k,n,y;x,t) &= \widehat{m}(k;x,t) m(n,y;k,x,t) \\ &= \frac{S_{kny}(x,t;\Theta)}{\sum\limits_{\forall k} \sum\limits_{\forall n} \sum\limits_{\forall y} S_{kny}(x,t;\Theta)} \end{aligned} \quad (30)$$

when $\Theta$ is completely fixed and then $\Theta$ should be updated using Eq. 29. Although this deterministic procedure is expected to give appropriate convergence of $\Theta$ when the initial point is chosen to be close to the global minimum, but may often fail if it is not. Ueda and Nakano (1998) considered that such common problem in EM algorithm is mainly due to the fact that the membership degree (missing data posterior density function) given by Eq. 30 is often unreliable at early stage of the iteration. They reformulated EM algorithm to improve its drawback, from the viewpoint of the statistical mechanics analogy. In place of $\widehat{m}(k,n,y;x,t)$, they gave the membership degree $\widehat{f}(k,n,y;x,t)$ parameterized by $\beta$ as

$$\widehat{f}(k,n,y;x,t,\beta) = \frac{S_{kny}(x,t;\Theta)^{\beta}}{\sum\limits_{\forall k} \sum\limits_{\forall n} \sum\limits_{\forall y} S_{kny}(x,t;\Theta)^{\beta}}. \quad (31)$$

Since in general cases, initial points are of course not always near the global solution, every cluster should share $W(x,t)$ almost evenly by setting $\beta \approx 0$ (where it contributes to smoothing J, that is perhaps often multimodal, i.e., 'spiky') in the early stage of the iteration and as the iteration proceeds, Eq. 31 should approach the original one (Eq. 30) by setting $\beta = 1$ to accomplish the primary objective. To achieve this, they newly added a $\beta$-loop in addition to the traditional EM loop. DAEM-based HTM optimization is implemented as following:

---

1. Set $\beta \leftarrow \beta_{\mathtt{min}} (0 < \beta_{\mathtt{min}} < 1)$
2. Set $\Theta^{(0)}$, $i \leftarrow 0$
3. Iterate EM-steps until convergence:
   E-step: Substitute $\Theta^{(i)}$ to Eq.31
   M-step: $\Theta^{(i+1)} \leftarrow$ Eq.29
   Set $i \leftarrow i+1$.

4. Increase $\beta$.
5. If $\beta < 1$, repeat from step 3; otherwise stop.

---

## 5 PARAMETER UPDATE EQUATIONS

For the purpose of reducing the dimension of the feature to extract, let us roughly assume that all pitch trajectories are parallel to the time axis (0-order polynomial), i.e.,

$$\mu_k(t) \approx \mu_{k0} \tag{32}$$

and each partial stream in a HTM has similar power envelope (only a single power envelope function is assumed in a single HTM so that the index $n$ in $U_{kn}(t)$ shall be excluded). Since our objective here is not to strictly analyze precise music expressions, these assumptions would not be fatal flaws in practical situation. From Eq. 21, logarithmic kernel density $\log S_{kny}(x,t;\Theta)$ is given by

$$\log S_{kny}(x,t;\Theta) = \log \frac{w_k v_{kn} u_{ky}}{2\pi\sigma_k\phi_k} - \frac{(x-\mu_{k0}-\log n)^2}{2\sigma_k^2} - \frac{(t-\tau_k-y\phi_k)^2}{2\phi_k^2} \tag{33}$$

so that solving Eq. 29, the update equation of each parameter at M-step is derived as follows:

$$\ell_{kny}(x,t;\beta) \equiv \widehat{f}(k,n,y;x,t,\beta)W(x,t)$$

$$\widehat{w}_k^{(i+1)} = \sum_{\forall n,\forall y}\iint_D \ell_{kny}(x,t;\beta)\,dx\,dt$$

$$\widehat{\mu}_{k0}^{(i+1)} = \frac{\sum_{\forall n,\forall y}\iint_D (x-\log n)\ell_{kny}(x,t;\beta)\,dx\,dt}{\widehat{w}_k^{(i+1)}}$$

$$\widehat{\tau}_k^{(i+1)} = \frac{\sum_{\forall n,\forall y}\iint_D (t-y\phi_k^{(i)})\ell_{kny}(x,t;\beta)\,dx\,dt}{\widehat{w}_k^{(i+1)}}$$

$$\widehat{v}_{kn}^{(i+1)} = \frac{d_v\bar{v}_n + \sum_{\forall y}\iint_D \ell_{kny}(x,t;\beta)\,dx\,dt}{d_v + \widehat{w}_k^{(i+1)}}$$

$$\widehat{u}_{ky}^{(i+1)} = \frac{d_u\bar{u}_n + \sum_{\forall n}\iint_D \ell_{kny}(x,t;\beta)\,dx\,dt}{d_u + \widehat{w}_k^{(i+1)}}$$

$$\widehat{\phi}_k^{(i+1)} = \frac{-\Lambda_k + \left(\Lambda_k^2 + 4\sum_{\forall y}\int \Upsilon_{ky}(t)^2(t-\tau_k)^2\,dt\right)^{\frac{1}{2}}}{2\widehat{w}_k^{(i+1)}}$$

$$\left(\begin{array}{l}\Lambda_k = \sum_{\forall n,\forall y}\iint_D y(t-\tau_k)\ell_{kny}(x,t;\beta)\,dx\,dt \\ \Upsilon_{ky}(t) = \sum_{\forall n}\int \ell_{kny}(x,t;\beta)\,dx\end{array}\right)$$

$$\widehat{\sigma}_k^{(i+1)} = \left(\frac{\sum_{\forall n,\forall y}\iint_D (x-\mu_{k0}^{(i)}-\log n)^2\ell_{kny}(x,t;\beta)\,dx\,dt}{\widehat{w}_k^{(i+1)}}\right)^{\frac{1}{2}}$$

Table 4: Experimental Conditions

| | | |
|---|---|---|
| frequency analysis | Sampling rate | 16 kHz |
| | frame shift | 16 ms |
| | frequency resolution | 12.0 cent |
| | frequency range | 60–3000 Hz |
| HTC | initial # of HTMs | 20 |
| | # of partials: $N$ | 6 |
| | # of kernels in $U_k(t)$: $Y$ | 10 |
| | $\beta_{min}$ | 0.5 |
| | $\bar{r}_n$ | $0.6547 \times n^{-2}$ |
| | $d_r$, $d_c$ | 0.04 |
| | time range of analyzing segment | 80 frames (1.28 s) |
| | # of analyzing segments | 21 (total time: 24 s) |
| PreFEst-core (Goto, 2004) | pitch resolution | 20 cent |
| | # of partials | 8 |
| | # of tone models | 200 |
| | standard deviation of Gaussian | 3.0 |
| | $\bar{r}_n$ | $0.6547 \times n^{-2}$ |
| | $\bar{d}$ (prior contribution factor) | 3.0 |

## 6 EXPERIMENTAL EVALUATION

### 6.1 CONDITIONS

To verify the potential performance of the proposed method as an audio feature extraction application, we tested it on a set of real performance music data excerpted from RWC music database (see table 2 for the list of the experimental data). Time series of power spectrum was analyzed using Gabor wavelet transform with frame shift of 16 ms for input digital signals of 16 kHz sampling rate. The lower bound of the frequency range and the frequency resolution were 60 Hz and 12 cent, respectively. The initial parameters of $(\mu_{k0},\tau_k|k=1,\cdots,K)$ for DAEM algorithm were automatically determined by picking 20 largest peaks in the observed spectral time pattern of 80 consecutive frames. After the parameters converged, the total number of note events were estimated by intensity thresholding, i.e., every HTM whose $w_k$ estimate becomes smaller than the threshold was truncated. See table 4 for the detailed conditions.

### 6.2 CALCULATING ACCURACY

Using the supplementary hand-labeled reference MIDI data, associated with each test data, the comprehensive accuracy of pitch, onset time and duration estimates was automatically calculated by the following procedure.

1. Truncate HTMs with intensity thresholding on $w_k$ estimate.
2. Quantize pitch estimate $\mu_{k0}$, onset time estimate $\tau_k$ and duration estimate $Y\phi_k$ to the closest note, frame and number of frames in each remaining HTM and then create a framewise binary series each for 128 number of notes, where 1 and 0 indicate 'note activation' and 'silence' at each frame, respectively.
3. Convert the hand-labeled reference MIDI data to a reference framewise binary series each for 128 number of notes similarly where 1 and 0 indicate 'note activation' and 'silence'.
4. Add up the accumulated costs, computed by Dynamic Programming (DP) matching between the two binary series, of 128 notes. Since the onset and offset times of respective note events in the real performance data and the reference MIDI data are not per-

Table 2: List of The Experimental Data Excerpted from RWC Music Database

| Symbol | Title (Genre) | Catalog number | Composer/Player | Instruments | # of frames |
|---|---|---|---|---|---|
| data(1) | Crescent Serenade (Jazz) | RWC-MDB-J-2001 No. 9 | S. Yamamoto | Guitar | 4427 |
| data(2) | For Two (Jazz) | RWC-MDB-J-2001 No. 7 | H. Chubachi | Guitar | 6555 |
| data(3) | Jive (Jazz) | RWC-MDB-J-2001 No. 1 | M. Nakamura | Piano | 5179 |
| data(4) | Lounge Away (Jazz) | RWC-MDB-J-2001 No. 8 | S. Yamamoto | Guitar | 9583 |
| data(5) | For Two (Jazz) | RWC-MDB-J-2001 No. 2 | M. Nakamura | Piano | 9091 |
| data(6) | Jive (Jazz) | RWC-MDB-J-2001 No. 6 | H. Chubachi | Guitar | 3690 |
| data(7) | Three Gimnopedies no. 1 (Classic) | RWC-MDB-C-2001 No. 35 | E. Satie | Piano | 6571 |
| data(8) | Nocturne no.2, op.9-2(Classic) | RWC-MDB-C-2001 No. 30 | F. F. Chopin | Piano | 7258 |

Table 3: Accuracy results of PreFEst-core (Goto, 2004) and HTC. Columns (A)~(J) and (K)~(R) show the accuracies with different thresholds for PreFEst-core and HTC, respectively: (A)$2.0\times10^8$, (B)$2.5\times10^8$, (C)$5.0\times10^8$, (D)$7.5\times10^8$, (E)$10\times10^8$, (F)$15\times10^8$, (G)$17.5\times10^8$, (H)$20\times10^8$, (I)$25\times10^8$, (J)$27.5\times10^8$, (K)$7.5\times10^9$, (L)$1.0\times10^{10}$, (M)$2.0\times10^{10}$, (N)$3.0\times10^{10}$, (O)$4.0\times10^{10}$, (P)$5.0\times10^{10}$, (Q)$6.0\times10^{10}$, (R)$7.0\times10^{10}$.

| | Accuracy(%) | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PreFEst-coreGoto (2004) | | | | | | | | | | HTC | | | | | | | |
| | (A) | (B) | (C) | (D) | (E) | (F) | (G) | (H) | (I) | (J) | (K) | (L) | (M) | (N) | (O) | (P) | (Q) | (R) |
| data(1) | 56.6 | 62.49 | 75.9 | 81.6 | 83.3 | **84.6** | 83.0 | 81.5 | 78.4 | 75.8 | 69.5 | 74.8 | 83.9 | 84.8 | 88.2 | **88.8** | 88.7 | 85.1 |
| data(2) | 68.7 | **69.6** | 66.3 | 59.0 | 53.7 | 36.3 | 32.4 | 30.3 | 26.8 | 26.5 | 84.3 | 88.2 | **90.6** | 82.5 | 75.7 | 72.3 | 67.9 | 61.9 |
| data(3) | -20.8 | -7.3 | 31.7 | 47.8 | 56.9 | 65.1 | 69.5 | 71.9 | **75.5** | 71.8 | 68.8 | 70.0 | 77.6 | 80.0 | **80.2** | 77.4 | 73.3 | 73.4 |
| data(4) | 55.1 | 56.8 | 60.7 | 63.3 | 63.1 | 63.6 | **64.1** | 62.3 | 60.6 | 60.2 | 82.6 | 83.0 | **83.8** | 82.4 | 82.8 | 82.0 | 81.5 | 76.5 |
| data(5) | 50.7 | 53.2 | **61.0** | 60.0 | 58.8 | 59.3 | 57.6 | 58.0 | 57.5 | 49.7 | 76.3 | 79.3 | 79.4 | **81.7** | 77.6 | 76.2 | 76.5 | 72.8 |
| data(6) | -7.2 | 6.6 | 37.9 | 51.1 | 57.7 | 65.9 | 65.6 | **66.7** | 66.3 | 65.7 | 77.5 | 79.6 | 81.7 | 82.7 | **84.4** | 82.3 | 81.4 | 80.7 |
| data(7) | 51.6 | 54.1 | **62.7** | 52.4 | 47.0 | 45.9 | 42.7 | 41.1 | 42.2 | 42.7 | **72.1** | 69.9 | 70.3 | 68.3 | 66.9 | 63.1 | 61.5 | 62.0 |
| data(8) | 20.8 | 22.9 | 36.6 | **42.5** | 38.5 | 39.1 | 38.8 | 37.7 | 32.7 | 30.6 | 73.7 | **75.9** | 75.6 | 72.2 | 67.6 | 61.1 | 48.9 | 46.7 |

fectly aligned, time warping technique was somehow required.

5. The accumulated cost divided by the total number of frames of note activation in 128 sets of reference binary series gives the error rate. The accuracy rate is simply given by subtracting the error rate from 1.

$$\text{Accuracy}(\%) = \frac{A - (\overbrace{\text{Ins} + \text{Del}}^{\text{accumulated cost}})}{A} \times 100$$

$A$ : total frame # of 'note activation'

Ins : # of insertion errors

Del : # of deletion errors

Note that this calculation counts a substitution error as duplicated errors (one deletion and one insertion errors), so that the accuracy can possibly take negative values.

### 6.3 RESULTS

We chose [1]'PreFEst-core'(Goto, 2004) for a comparison, as it is recently accepted as one of the most successful methods developed for multi-pitch analysis. Since PreFEst-core extracts the most dominant pitch trajectory from multi-pitch signals and does not include a specific procedure of estimating the number of sources, we included intensity thresholding similarly for pitch candidate truncation. A typical example of the estimated binary series extracted via step 2 mentioned in 6.2 on particular test data is shown in Fig.5 together with the hand-labeled reference MIDI data. The optimized model and the corresponding observed spectral time pattern are shown with 3D and grayscale displays in Fig.4.

[1]Note that we have only implemented the 'PreFEst-core', i.e., a framewise pitch likelihood estimation, for the evaluation and not included the 'PreFEst-back-end', i.e., multi-agent based pitch tracking algorithm.

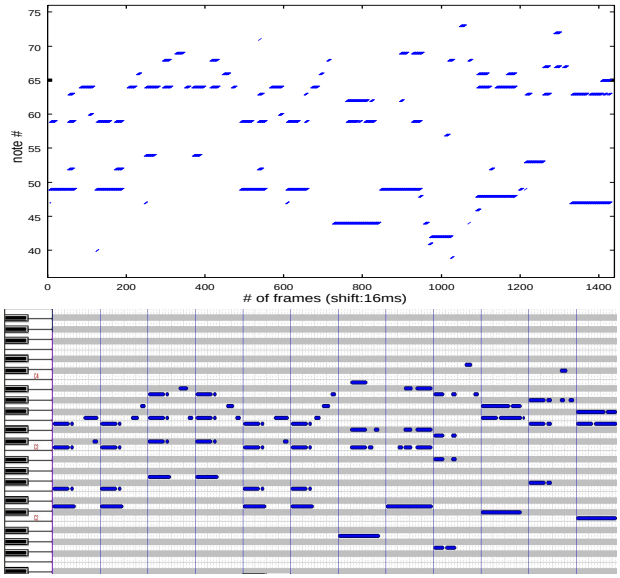
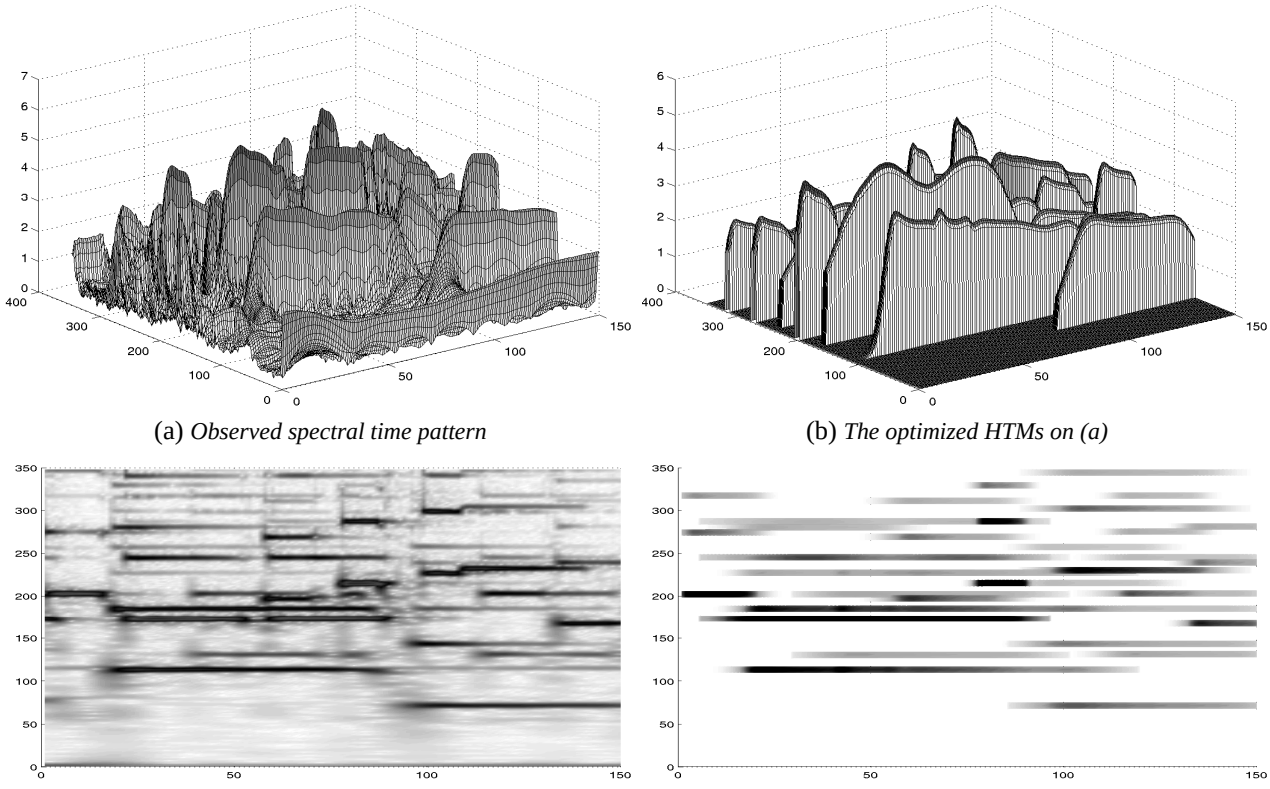



Figure 5: Estimates of $\mu_{k0}, \tau_k, Y\phi_k$(top) and the reference MIDI data displayed in piano-roll form (bottom).

In thresholding, there is a trade-off between the number of insertion and deletion errors according to the threshold degree. Therefore, to properly validate the performance of the proposed method, we considered that the maximal accuracy among all the thresholds that were tested, which might show the limit of a potential capability, should be a criterion for comparison. Accuracy results of PreFEst-core and HTC with different degrees of truncation thresholding are shown in table 3. The number in bold-faced type is the best accuracy in each data among all the thresholds, which we are only concerned with. Comparing these accuracies between PreFEst-core and HTC, HTC significantly outperforms PreFEst-core, from which its potential is clearly verified.



(a) *Observed spectral time pattern*

(b) *The optimized HTMs on (a)*

(c) *spectrogram of the observation*

(d) *A grayscale display of the optimized model*

Figure 4: 3D and grayscale diplays of the given spectrum and the parameter-optimized model

## 7 CONCLUSION

We established a new framework for multi-pitch analysis based upon two dimensional geometric modeling and estimation of the distinct spectral streams localized in time-frequency 'scene', and investigated possibilities for an application of audio feature extraction available for MIR.

The method described in this paper still has many interesting issues to consider, e.g., estimation of the number of note events without relying on heuristic thresholding, further precise modeling by introducing higher order coefficients in pitch trajectory polynomial and inharmonicity factor parameter and others for sound segregation or noise reduction applications.

## REFERENCES


M. Abe and S. Ando. Auditory scene analysis based on time-frequency integration of shared fm and am (ii): Optimum time-domain integration and stream sound reconstruction. *IEICE Trans.*, J83-D-II(2):468–477, 2000. in Japanese.

D. Chazan, Y. Stettiner, and D. Malah. Optimal multi-pitch estimation using the em algorithm for co-channel speech separation. In *Proc. ICASSP'93*, volume 2, pages 728–731, 1993.

M. Feder and E. Weinstein. Parameter estimation of superimposed signals using the em algorithm. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-36(4): 477–489, 1988.

S. Godsill and M. Davy. Baysian harmonic models for musical pitch estimation and analysis. In *Proc. ICASSP2002*, volume 2, pages 1769–1772, 2002.

M. Goto. A real-time music-scene-description system: Predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *ISCA Journal*, 43(4):311–329, 2004.

H. Kameoka, T. Nishimoto, and S. Sagayama. Minimum bic estimation of harmonic kernel regression model for multi-pitch analysis. *IEEE Trans. Speech and Audio Processing*, 2005. submitted.

K. Kashino, K. Nakadai, and H. Tanaka. Organization of hierarchical perceptual sounds: Music scene analysis with autonomous processing modules and a quantitive information integration mechanism. In *Proc. IJCAI*, volume 1, pages 158–164, 1995.

A. Klapuri, T. Virtanen, and J. Holm. Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals. In *Proc. COST-G6 Conference on Digital Audio Effects*, pages 233–236, 2000.

T. Nakatani. Computational auditory scene analysis based on residue-driven architecture and its application to mixed speech recognition. *Ph.D. thesis, Kyoto University*, 2002.

K. Nishi, S. Ando, and S. Aida. Optimum harmonics tracking filter for auditory scene analysis. In *Proc. ICASSP'96*, pages 573–576, 1996.

N. Ueda and R. Nakano. Deterministic annealing em algorithm. *Neural Networks*, 11(2):271–282, 1998.