A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering

Hirokazu Kameoka, *Student Member, IEEE*, Takuya Nishimoto, and Shigeki Sagayama, *Member, IEEE*

***Abstract*—This paper proposes a multipitch analyzer called the harmonic temporal structured clustering (HTC) method, that jointly estimates pitch, intensity, onset, duration, etc., of each underlying source in a multipitch audio signal. HTC decomposes the energy patterns diffused in time-frequency space, i.e., the power spectrum time series, into distinct clusters such that each has originated from a single source. The problem is equivalent to approximating the observed power spectrum time series by superimposed HTC source models, whose parameters are associated with the acoustic features that we wish to extract. The update equations of the HTC are explicitly derived by formulating the HTC source model with a Gaussian kernel representation. We verified through experiments the potential of the HTC method.**

***Index Terms*—Computational acoustic scene analysis, harmonic temporal structured clustering (HTC), multipitch analyzer.**

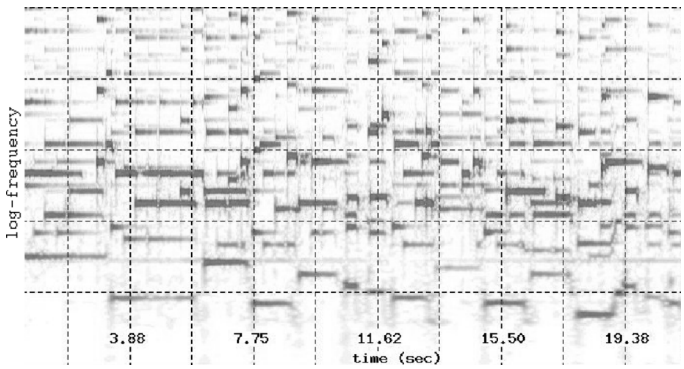


Fig. 1. Wavelet spectrogram of real performed music signal ranging from T_0 to T_1 in time direction and from X_0 to X_1 in frequency direction.

I. Introduction

WE HAVE been working on a new method for computational acoustic scene analysis having in mind, for example, a music content description system, a new equalizer system enabling volume and bass/treble controls for each separate sound source, a music information retrieval (MIR) system, and a precise acoustic environment [background music (BGM), phone ringing, interfering speech, etc.] detector for a wide range of speech applications. This paper describes a new multipitch analyzer that jointly estimate acoustic features such as pitch, onset, duration, energy, spectral, and temporal envelopes of each underlying source in a multipitch acoustic signal.

While the standard level of the numerous conventional multipitch analyzers has been considered to be far from practical use, recent pioneering ideas, e.g., graphical model-based [1], filterbank-based [2], nonparametric Kalman filtering-based [3], [4], multiagent-based [5], cochlear filtering-based [6], and parametric signal and spectrum modelings-based approaches [7]–[11] have brought remarkable progress to the practical step. Most of these methods take two major steps to resolve the problem: they start with a separation or a pitch feature extraction of concurrent sources at each short-time segment (frame) and then find the most likely overall pitch trajectories along time. In auditory scene analysis (ASA), these two processes in human audition are generally called "segregation" and "integration," respectively.

It is quite obvious that the more accurate the segregation process, the more reliable the result of the integration process. On the other hand, we hope to know, if possible, the pitches and the spectral component powers at preceding and succeeding frames to estimate a high-precision result of the segregation process at the current frame assuming they change gradually over time. Therefore, these two processes should be done essentially in a cooperative way and not independently with successive estimations for even more reliable results. This belief has led us to formulate a unified estimation framework for the two-dimensional structure of time-frequency power spectra, in contrast to the conventional strategy. The method that is presented in this article formulates the problem as a localization and shape detection of distinct spectrogram portions in the acoustic scene (time-frequency space), which essentially amounts to the estimation of pitch contour, onset, duration, and timbre feature of each sound source.

II. Formulation

Consider an observed power spectrum time series $W(x,t)$ (see Fig. 1), where x and t are log-frequency and time, defined on a domain of definition

$$D = \{x,t \in \mathbb{R} \mid X_0 \leqq x \leqq X_1, T_0 \leqq t \leqq T_1\}. \quad (1)$$

The problem of interest is to decompose this observed pattern into K sequential spectral streams, i.e., clusters, each of which is assumed to have originated from a single distinct source activation. This problem is an unsupervised categorization of the energy density $W(x,t)$ at each coordinate (x,t).

Manuscript received May 6, 2005; revised July 11, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael Davies.

The authors are with Graduate School of Information Science and Technology, University of Tokyo, Tokyo 113-8656, Japan (e-mail: kameoka@hil.t.u-tokyo.ac.jp; nishi@hil.t.u-tokyo.ac.jp; sagayama@hil.t.u-tokyo.ac.jp).

Color versions of Figs. 5–9 are available online at http://ieeexplore.ieee.org.
Digital Object Identifier 10.1109/TASL.2006.885248

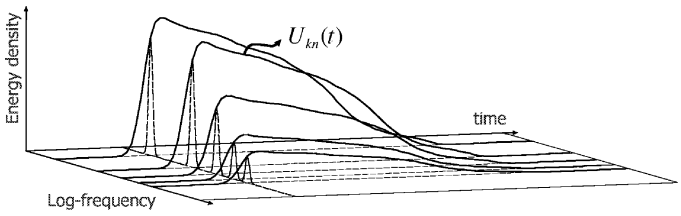


Fig. 2. Profile of the kth HTC source model $q_k(x,t;\mathbf{\Theta})$ ((19)).

The observed energy density $W(x,t)$ at each coordinate (x,t) has not necessarily originated from a single component in a single source but is in the general case rather a mixture of the energy patterns generated from other components or different sources. Thus, it is natural to assume that the energy density $W(x,t)$ has an unknown fuzzy membership to the kth source, introduced as a spectral masking function $m_k(x,t)$. Approximately assuming that observed power spectral densities are the sum of the independent power densities generated from underlying sources, which is true in the expectation sense (if we assume the phase of every component is uniformly random), $m_k(x,t)$ must satisfy

$$\forall x, \forall t : \sum_k m_k(x,t) = 1,\ \forall k : 0 < m_k(x,t) < 1. \quad (2)$$

Therefore, $m_k(x,t)W(x,t)$ denotes the decomposed spectral density of the kth source, which we will refer to as the kth cluster. We may wish to decompose $W(x,t)$ such that all clusters are temporally continuous and also have a harmonic structure, if we assume all sources are periodic signals and their components evolve gradually over time. Let us define here $q_k(x,t;\mathbf{\Theta})$, governed by parameter vector $\mathbf{\Theta}$, that characterizes the typical spectrogram of a single source (a graphical representation can be seen in Fig. 2). Note that the class of the sources mentioned here, in general, includes not only periodic signals but also nonperiodic signals such as white/pink noises, drum sounds or any others, as far as those spectrograms can be modeled with a suitable expression. Using $q_k(x,t;\mathbf{\Theta})$, we are now able to assess a "goodness" of the partitioned cluster $m_k(x,t)W(x,t)$ with the Kullback–Leibler (KL) divergence of $m_k(x,t)W(x,t)$ and $q_k(x,t;\mathbf{\Theta})$

$$\iint_D m_k(x,t)W(x,t)\log\frac{m_k(x,t)W(x,t)}{q_k(x,t;\mathbf{\Theta})}\mathrm{d}x\mathrm{d}t \quad (3)$$

with

$$\iint_D m_k(x,t)W(x,t)\mathrm{d}x\mathrm{d}t = \iint_D q_k(x,t;\mathbf{\Theta})\mathrm{d}x\mathrm{d}t. \quad (4)$$

The condition given as (4) makes the distortion measure (3) nonnegative. In the case of discrete-time observations, the distortion measure is still given by (3) where the integral is replaced by the sum over all discrete points of x and t. Now one notices that as $q_k(x,t;\mathbf{\Theta})$ and $m_k(x,t)W(x,t)$ become closer, (3) approaches zero. Hence, one can choose as the global cost function of the clustering to minimize w.r.t. $m_k(x,t)$ and $\mathbf{\Theta}$, the sum over k of the above measure

$$J \triangleq \sum_{k=1}^{K}\iint_D m_k(x,t)W(x,t)\log\frac{m_k(x,t)W(x,t)}{q_k(x,t;\mathbf{\Theta})}\mathrm{d}x\mathrm{d}t. \quad (5)$$

The unknown variables being m and $\mathbf{\Theta}$, the optimization we are solving is summarized as

$$\{\widehat{m},\widehat{\mathbf{\Theta}}\} = \operatorname*{argmin}_{m,\mathbf{\Theta}} J$$
$$m = \left\{m_k(x,t)\ \middle|\ k = 1,\ldots,K,\ x,t\in\mathbb{R}\right\}. \quad (6)$$

In order to find the optimal m and $\mathbf{\Theta}$, we shall find it most convenient to recursively optimize m and $\mathbf{\Theta}$ while keeping the other variable fixed as there is no analytical solution to (6). As each iteration necessarily decreases J, which is bounded below, m and $\mathbf{\Theta}$ gradually converge to a stationary point. Note, however, that with this procedure m and $\mathbf{\Theta}$ do not always converge to a global optimum but possibly to a local optimum.

The optimal m when $\mathbf{\Theta}$ is fixed can be obtained by finding the extreme value of

$$J - \iint_D \lambda(x,t)\left(\sum_k m_k(x,t) - 1\right)\mathrm{d}x\mathrm{d}t \quad (7)$$

where λ is a Lagrange undetermined multiplier. The partial derivative of the integrand w.r.t. $m_k(x,t)$ is thus given as

$$W(x,t)\left(1+\log\frac{m_k(x,t)W(x,t)}{q_k(x,t;\mathbf{\Theta})}\right) - \lambda(x,t). \quad (8)$$

Setting it to zero, one obtains

$$m_k(x,t) = \frac{q_k(x,t;\mathbf{\Theta})}{W(x,t)}\exp\left(\frac{\lambda(x,t)}{W(x,t)} - 1\right). \quad (9)$$

From (2) and (9), we have

$$\lambda(x,t) = W(x,t)\left(1 - \log\frac{\sum_k q_k(x,t;\mathbf{\Theta})}{W(x,t)}\right). \quad (10)$$

Substituting (10) in (9), one finally obtains the optimal spectral masking function for a fixed $\mathbf{\Theta}$ as

$$\widehat{m}_k(x,t) = \frac{q_k(x,t;\mathbf{\Theta})}{\sum_k q_k(x,t;\mathbf{\Theta})}. \quad (11)$$

Substituting (11) in (5), it becomes clear that during this clustering we are also iteratively decreasing the KL divergence between the whole observed spectrogram $W(x,t)$ and the sum of $q_k(x,t;\mathbf{\Theta})$ over k

$$\min_m J = \iint_D W(x,t)\log\frac{W(x,t)}{\sum_k q_k(x,t;\mathbf{\Theta})}\mathrm{d}x\mathrm{d}t. \quad (12)$$

Therefore, this clustering is understood as a geometric optimal approximation of $W(x,t)$ using the model $q_k(x,t;\mathbf{\Theta})$.

Note that this result proves without using Bayes rules the convergence of the expectation-maximization (EM) algorithm, as one regards k as missing data and $q_k(x,t;\boldsymbol{\Theta})$ as a complete data pdf $p(k,x,t|\boldsymbol{\Theta})$. The correspondence to the EM algorithm becomes much clearer by comparing the result of (11) and the terms in J that depends on $\boldsymbol{\Theta}$

$$\widetilde{J} = \sum_k \iint_D m_k(x,t) W(x,t) \log q_k(x,t;\boldsymbol{\Theta}) \mathrm{d}x\mathrm{d}t \tag{13}$$

with the Q function, given by

$$\begin{aligned} &Q(\boldsymbol{\Theta}, \widetilde{\boldsymbol{\Theta}}) \\ &= \sum_k \iint_D \overbrace{p(k|x,t,\boldsymbol{\Theta})}^{\text{missing data pdf}} \overbrace{W(x,t)}^{\text{observed pdf}} \log \overbrace{p(k,x,t|\widetilde{\boldsymbol{\Theta}})}^{\text{complete data pdf}} \mathrm{d}x\mathrm{d}t \\ &p(k|x,t,\boldsymbol{\Theta}) \\ &= \frac{p(k,x,t|\boldsymbol{\Theta})}{p(x,t|\boldsymbol{\Theta})} = \frac{p(k,x,t|\boldsymbol{\Theta})}{\sum_k p(k,x,t|\boldsymbol{\Theta})}. \end{aligned}$$

This discussion implies that practically the same procedure as the EM algorithm can be used even though $W(x,t)$ and $q_k(x,t;\boldsymbol{\Theta})$ are not pdf's.

With fixed membership degree $m_k(x,t)$ updated by the E-step (11), the parameter $\boldsymbol{\Theta}$, on the other hand, should be updated by the M-step

$$\widehat{\boldsymbol{\Theta}} = \operatorname*{argmin}_{\boldsymbol{\Theta}} J \tag{14}$$

depending on the specific form of $q_k(x,t;\boldsymbol{\Theta})$. The update equation for $\boldsymbol{\Theta}$ will be given in Section V after $q_k(x,t;\boldsymbol{\Theta})$ is introduced. We will call this approach "harmonic temporal structured clustering (HTC)" in the particular case where $q_k(x,t;\boldsymbol{\Theta})$ is assumed to have a harmonic structure.

The HTC procedure is summarized as follows. The input to the system is an observed (known) signal, characterized by its spectrogram $W(x,t)$, where x and t are log-frequency and time. The membership degree $m_k(x,t)$ of kth source/stream is unknown (spectral masking function). On the other hand, the spectrogram of the kth source can be modeled by a function $q_k(x,t;\boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ is the set of model parameters. These are the unknown variables that we want to estimate. The HTC method works by iteratively updating the estimates of: 1) $m_k(x,t)$ with $\boldsymbol{\Theta}$ fixed by (11) and 2) $\boldsymbol{\Theta}$ with $m_k(x,t)$ fixed.

III. HTC Model

A. Model Representation

In this section, we will introduce the HTC source model $q_k(x,t;\boldsymbol{\Theta})$. Let us assume through the rest of this paper that all sources are periodic signals having smooth power envelopes. Supposing the pitch contour during a single source activation is expressed with a polynomial (imagine vibrato or glissando)

$$\mu_k(t) = \mu_{k0} + \mu_{k1} t + \mu_{k2} t^2 + \ldots \tag{15}$$

a cutting plane of $q_k(x,t;\boldsymbol{\Theta})$ at particular time t represents a harmonic structure of pitch frequency $\mu_k(t)$ (see Fig. 3).

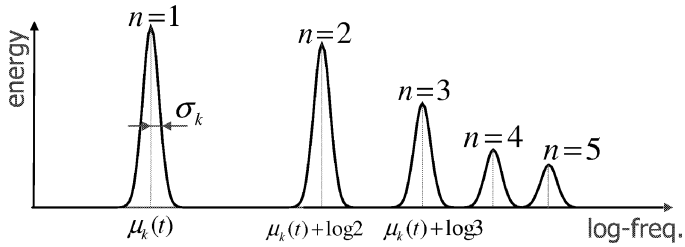


Fig. 3. Cutting plane of $q_k(x,t;\boldsymbol{\Theta})$ at time t.

Given the pitch contour $\mu_k(t)$ in kth HTC source model, the contour of the nth partial is $\mu_k(t) + \log n$. Now let the frequency spread of each harmonic component be approximated by a Gaussian distribution function when the spectra are obtained by[1] the wavelet transform (constant Q transform) using Gabor-wavelet basis function. The accuracy of this approximation is investigated through a numerical computation in Appendix I. Denoting by $U_{k,n}(t)$ the power envelope of the nth partial (presumed to be a function that is normalizable since $q_k(x,t;\boldsymbol{\Theta})$ has to satisfy (4)), such that

$$\forall k, \forall n : \int_{-\infty}^{\infty} U_{k,n}(t) \mathrm{d}t = 1 \tag{16}$$

then the normalized energy density of the nth partial in the kth HTC source model is given as a multiplication of $U_{k,n}(t)$ and a Gaussian distribution centered at $\mu_k(t) + \log n$

$$U_{k,n}(t) \times \frac{v_{k,n}}{\sqrt{2\pi}\sigma_k} e^{-(x-\mu_k(t)-\log n)^2/2{\sigma_k}^2}, \qquad n = 1, \ldots, N \tag{17}$$

where σ_k denotes the frequency spread of every partial, and $v_{k,n}$ is the relative energy of the nth partial, satisfying

$$\forall k : \sum_n v_{k,n} = 1. \tag{18}$$

Therefore, the power density of the kth HTC source model $q_k(x,t;\boldsymbol{\Theta})$ as a whole (see Fig. 2) can be written as

$$q_k(x,t;\boldsymbol{\Theta}) = w_k \sum_{n=1}^{N} \frac{v_{k,n} U_{k,n}(t)}{\sqrt{2\pi}\sigma_k} e^{-(x-\mu_k(t)-\log n)^2/2{\sigma_k}^2} \tag{19}$$

where w_k indicates the total energy of the kth source. The sum of a number K of the HTC source models $\sum_{k=1}^{K} q_k(x,t;\boldsymbol{\Theta})$ is supposed to be a model for the observed overall spectrogram $W(x,t)$.

Let us now discuss how we should model the power envelope function $U_{k,n}(t)$. In the general case where one does not know in advance what the sources are, it is perhaps wise not to use a physically oriented model for $U_{k,n}(t)$ based only upon a particular sound production mechanism. It is thus important to introduce for $U_{k,n}(t)$ as generic a model as possible.

What we hope to choose for $U_{k,n}(t)$ is a function that is temporally continuous, nonnegative, having a time spread from

[1]Having music applications in mind, since usually the musical notes are equally spaced on a log-frequency scale, there is a need to use a transform that has a variable window length characteristic. In particular, it is difficult to distinguish the sources of lower notes with short-time Fourier transform spectra, for example, that has a fixed window length across all frequencies. This is one of the reasons we chose to use wavelet transform as a front-end for HTC.

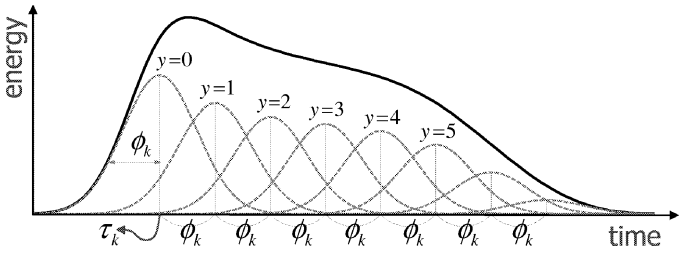


Fig. 4. Power envelope function $U_{k,n}(t)$ [(20)].

minus to plus infinity (assuming the Gabor-wavelet basis as the mother wavelet) and adaptable to various curves. Furthermore, it should satisfy (16). We came up with a function satisfying all these requirements, given as

$$U_{k,n}(t) = \sum_{y=0}^{Y-1} \frac{u_{k,n,y}}{\sqrt{2\pi}\phi_{k,n}} \exp\left(-\frac{(t-\tau_k - y\phi_{k,n})^2}{2{\phi_{k,n}}^2}\right). \quad (20)$$

τ_k is the center of the forefront Gaussian, that could be considered as an onset time estimate, $u_{k,n,y}$ the weight parameter for each kernel, that allows the function to have variable shapes. To satisfy (16), $u_{k,n,y}$ must only be normalized to unity

$$\forall k, \forall n : \sum_y u_{k,n,y} = 1. \quad (21)$$

The particularity of this function is that the centers of the Gaussian function kernels are spaced by a distance proportional to the common diffusion parameter $\phi_{k,n}$ with a proportionality coefficient α, which we henceforth set to 1 (see Fig. 4). This tying ensures the smoothness of the curve by preventing adjacent kernels to be separated from each other. $\phi_{k,n}$ also works as a parameter to make a linear stretch of $U_{k,n}(t)$ in the time direction allowing to express various durations of sources. Moreover, by forbidding switches in the position of the kernels, it reduces the singularity of the system, improving the optimization perspectives.

All the parameters of the HTC model are listed in Table I.

B. Subclustering

While the representation of $q_k(x,t;\mathbf{\Theta})$, the HTC source model, has been introduced, we are not yet able to obtain the analytical solution for (14) for each parameter in $\mathbf{\Theta}$. However, as the HTC source model is specified as the sum of the subsource model $S_{k,n,y}(x,t;\mathbf{\Theta})$

$$q_k(x,t;\mathbf{\Theta}) = \sum_n \sum_y S_{k,n,y}(x,t;\mathbf{\Theta}) \quad (22)$$

where

$$S_{k,n,y}(x,t;\mathbf{\Theta}) \triangleq \frac{w_k v_{k,n} u_{k,n,y}}{2\pi\sigma_k\phi_{k,n}} e^{(-(x-\mu_k(t)-\log n)^2/2\sigma_k^2)-(t-\tau_k-y\phi_{k,n})^2/2\phi_{k,n}^2} \quad (23)$$

the problem could be equivalently simplified by further breaking each cluster down into $\{n,y\}$-labeled subclusters, associated with the subsource model $S_{k,n,y}(x,t;\mathbf{\Theta})$.

Introducing now another masking function $m_{k,n,y}(x,t)$ that decomposes the kth partitioned cluster $m_k(x,t)W(x,t)$ into the $\{n,y\}$th subcluster and satisfies for all x and t

$$\forall k : \sum_n \sum_y m_{k,n,y}(x,t) = 1$$
$$\forall n, \forall y : 0 < m_{k,n,y}(x,t) < 1$$

we have the Jensen's inequality for all k, as shown in (24) at the bottom of the page, and equality holds when

$$m_{k,n,y}(x,t) = \frac{S_{k,n,y}(x,t;\mathbf{\Theta})}{\sum_n \sum_y S_{k,n,y}(x,t;\mathbf{\Theta})}. \quad (25)$$

The proof is omitted since it can be easily obtained by following the same way as in Section II. This means that if and only if $m_{k,n,y}(x,t)$ is given by (25), $J^+ = \sum_k J_k^+$ is minimized to the global cost function $J = \sum_k J_k$. One notices that obtaining parameter update equations through J^+ seems much easier than (14) since the summation over k, n, and y no longer appears inside the logarithm function. After making $J = J^+$ with the update of $m_{k,n,y}(x,t)$ by (25), one can indirectly decrease the objective function J by decreasing J^+ through an update of $\mathbf{\Theta}$. This is because J is always guaranteed by the inequation (24) to be even smaller than J^+. Decreasing J^+ w.r.t. $\mathbf{\Theta}$ can be done by increasing (26), as shown at the bottom of the next page.

The HTC procedure in the particular case where the HTC source model is given by (22) is again summarized as follows. The masking function $m_k(x,t)m_{k,n,y}(x,t)$, among the total energy density $W(x,t)$ for each subcluster is unknown. On the other hand, the HTC subsource model for the subcluster energy density is $S_{k,n,y}(x,t;\mathbf{\Theta})$, where $\mathbf{\Theta}$ is the set of model parameters (pitch, spectral envelope, temporal envelope, intensity, harmonicity, cf. Table I). The HTC works by iteratively updating the estimates of

1) (E-step) $m_k(x,t)m_{k,n,y}(x,t)$ by

$$m_k(x,t)m_{k,n,y}(x,t) = \frac{S_{k,n,y}(x,t;\mathbf{\Theta})}{\sum_k \sum_n \sum_y S_{k,n,y}(x,t;\mathbf{\Theta})} \quad (27)$$

2) (M-step) $\mathbf{\Theta}$ using $\widetilde{J}^+$.

$$J_k \triangleq \iint_D m_k(x,t)W(x,t)\log\frac{m_k(x,t)W(x,t)}{\sum_{n,y} S_{k,n,y}(x,t;\mathbf{\Theta})}\mathrm{d}x\mathrm{d}t$$
$$\leqq J_k^+ \triangleq \sum_{n,y}\iint_D m_k(x,t)m_{k,n,y}(x,t)W(x,t)\log\frac{m_k(x,t)m_{k,n,y}(x,t)W(x,t)}{S_{k,n,y}(x,t;\mathbf{\Theta})}\mathrm{d}x\mathrm{d}t \quad (24)$$

TABLE I
PARAMETERS OF THE HTC SOURCE MODEL

denotation	physical meanings
$\mu_k(t)$	pitch contour of the source (0-order polynomial would be a reasonable way to use)
w_k	total energy of the source
$v_{k,n}$	relative energy of n-th harmonic (perhaps useful as a timbre feature)
$u_{k,n,y}$	coefficient of the power envelope function of n-th partial
τ_k	onset time
$Y\phi_k$	duration (Y is a constant)
σ_k	diffusion in the frequency direction of the harmonics

With this procedure, J converges to a stationary point as well as with the procedure described in Section III.

The philosophy we adopt in the HTC is to approximate a source spectrogram in a real environment as closely as possible by a mathematically simple and compact model characterized by meaningful parameters rather than to try to make a perfect fit by an extremely complex model whose parameters are no longer meaningful.

IV. EXTENSION TO MAP AND THE USE OF PRIORS

Keeping only the terms depending on $\boldsymbol{\Theta}$ in (12) and taking the opposite, one defines the following function to maximize w.r.t. $\boldsymbol{\Theta}$:

$$\iint_D W(x,t)\log\sum_k q_k(x,t;\boldsymbol{\Theta})\mathrm{d}x\mathrm{d}t. \tag{28}$$

Using this function and letting Ω be

$$Q=\iint_D\sum_k q_k(x,t;\boldsymbol{\Theta})\mathrm{d}x\mathrm{d}t \tag{29}$$

one can derive the likelihood of the parameter $\boldsymbol{\Theta}$

$$P(W|\boldsymbol{\Theta})\triangleq\delta\left(\iint_D W(x,t)\mathrm{d}x\mathrm{d}t-\Omega\right) C\exp\left(\iint_D W(x,t)\log\sum_k q_k(x,t;\boldsymbol{\Theta})\mathrm{d}x\mathrm{d}t\right) \tag{30}$$

where Dirac delta $\delta(\cdot)$ ensures that the density is zero if

$$\iint_D W(x,t)\mathrm{d}x\mathrm{d}t\neq\Omega. \tag{31}$$

The parameter C given by

$$C=\frac{\Gamma(\Omega+1)}{\Omega^\Omega}\exp\left(-\iint_D\Gamma\big(W(x,t)+1\big)\mathrm{d}x\mathrm{d}t\right) \tag{32}$$

ensures that we obtain a probability measure where $\Gamma(\cdot)$ is the Gamma function. One can indeed see this probability as the joint probability of all the variables $W(x,t)$ following a multinomial-like distribution of parameter $\sum_k q_k(x,t;\boldsymbol{\Theta})$. This way of presenting the problem such that

$$\operatorname*{argmin}_{\boldsymbol{\Theta}}\min_{m_k(x,t)}J=\operatorname*{argmax}_{\boldsymbol{\Theta}}P(W|\boldsymbol{\Theta}) \tag{33}$$

enables us to extend it as a maximum *a posteriori* (MAP) estimation problem and to introduce prior distributions on the parameters as follows, using Bayes theorem

$$\begin{aligned}\widehat{\boldsymbol{\Theta}}&=\operatorname*{argmax}_{\boldsymbol{\Theta}}P(\boldsymbol{\Theta}|W)\\&=\operatorname*{argmax}_{\boldsymbol{\Theta}}\Big(\log P(W|\boldsymbol{\Theta})+\log P(\boldsymbol{\Theta})\Big)\\&=\operatorname*{argmin}_{\boldsymbol{\Theta}}\left(\min_{m_k(x,t)}J-\log P(\boldsymbol{\Theta})\right)\\&=\operatorname*{argmin}_{\boldsymbol{\Theta}}\left(\min_{m_k(x,t),m_{k,n,y}(x,t)}J^+-\log P(\boldsymbol{\Theta})\right).\end{aligned} \tag{34}$$

This optimization can be performed by iteratively updating $m_k(x,t)m_{k,n,y}(x,t)$ by (27) and $\boldsymbol{\Theta}$ using $\widetilde{J}^+ + \log P(\boldsymbol{\Theta})$.

Prior distribution works as a penalty function that tries to keep the parameters within a specified range. By introducing such a prior distribution on $v_{k,n}$, it becomes possible to prevent subharmonic (half-pitch) errors, as the resulting source model would usually have a harmonic structure with zero power for all the odd order harmonics, which is abnormal for speech and for many instruments. A prior distribution on $u_{k,n,y}$, on the other hand, helps to avoid overfitting many source models to the observed power envelope of a single source, as the resulting individual source models in this case would often have abnormal power envelopes.

For this purpose, we apply the Dirichlet distribution as the prior distribution, which is explicitly given by

$$P(\boldsymbol{v}_k)\triangleq\frac{\Gamma\Big(\sum_n(d_v\bar{v}_n+1)\Big)}{\prod_n\Gamma(d_v\bar{v}_n+1)}\prod_n {v_{k,n}}^{d_v\bar{v}_n} \tag{35}$$

$$P(\boldsymbol{u}_{k,n})\triangleq\frac{\Gamma\Big(\sum_y(d_u\bar{u}_y+1)\Big)}{\prod_y\Gamma(d_u\bar{u}_y+1)}\prod_y {u_{k,n,y}}^{d_u\bar{u}_y} \tag{36}$$

$$\widetilde{J}^+\triangleq\sum_k\sum_{n,y}\iint_D m_k(x,t)m_{k,n,y}(x,t)W(x,t)\log S_{k,n,y}(x,t;\boldsymbol{\Theta})\mathrm{d}x\mathrm{d}t \tag{26}$$

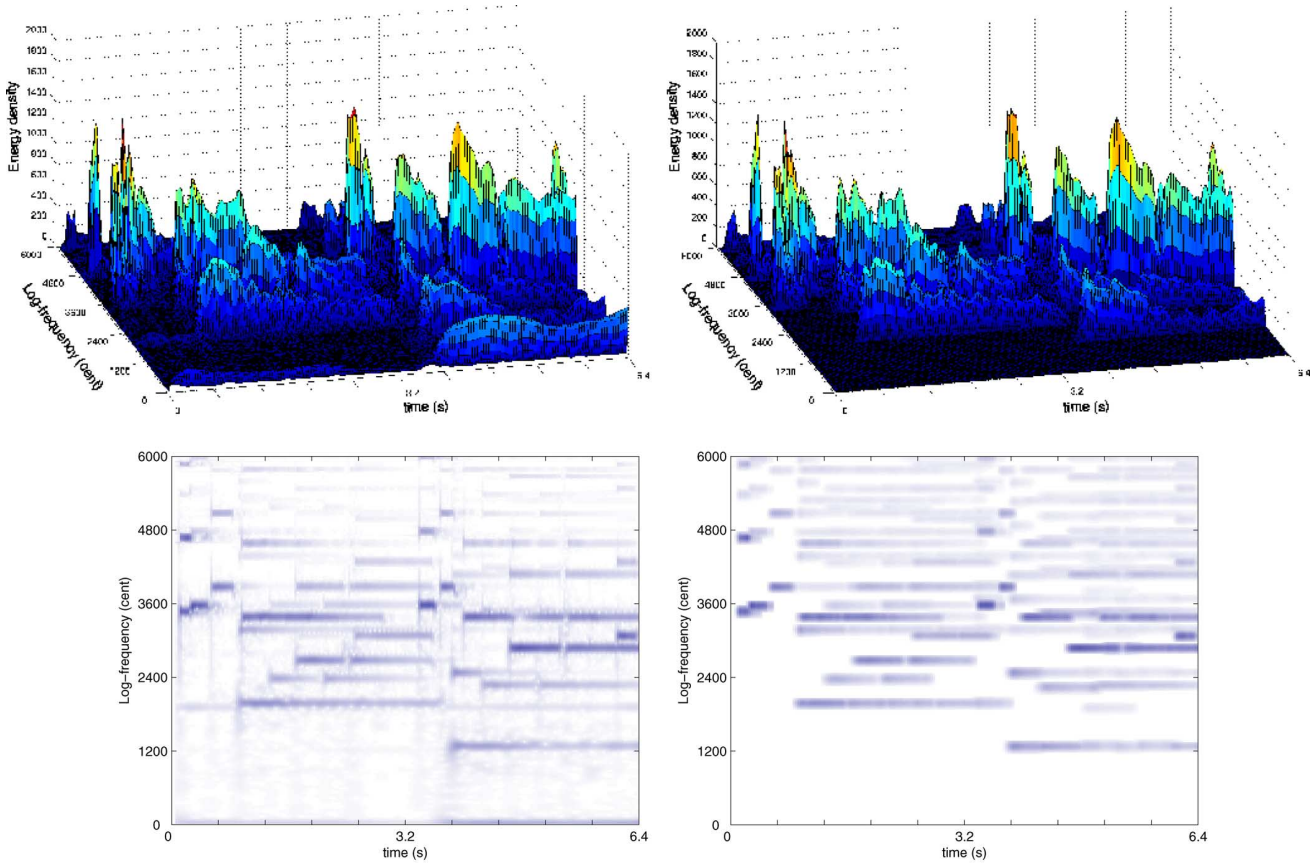


Fig. 5. Three-dimensional and top views of the observed power spectrum time series (top and bottom left) and the optimized HTC source models (top and bottom right).

where $\bar{v}_n$ and $\bar{u}_y$ are the most preferred "expected" values of $v_{k,n}$ and $u_{k,n,y}$ such that $\sum_n \bar{v}_n = 1$ and $\sum_y \bar{u}_y = 1$, d_v, and d_u regulate the strength of the priors. When d_v and d_u are zero, $P(\boldsymbol{v}_k)$ and $P(\boldsymbol{u}_{k,n})$ become uniform distributions. The maximum value for $P(\boldsymbol{v}_k)$ and $P(\boldsymbol{u}_{k,n})$ are taken respectively when $v_{k,n} = \bar{v}_n$ and $u_{k,n,y} = \bar{u}_y$ for all n and y if $d_v \neq 0$ and $d_u \neq 0$. The choice of this particular distribution allows us to give an analytical form of the update equations of $v_{k,n}$ and $u_{k,n,y}$. The joint prior distribution is now given by

$$P(\boldsymbol{\Theta}) = \prod_k P(\boldsymbol{v}_k) \prod_n P(\boldsymbol{u}_{k,n}). \tag{37}$$

Denoting by γ_w, γ_v^k and $\gamma_u^{k,n}$ the Lagrange multipliers for w_k, $v_{k,n}$ and $u_{k,n,y}$, respectively, the update equation of $\boldsymbol{\Theta}$ (M-step) that is guaranteed to increase $\widetilde{J}^+ + \log P(\boldsymbol{\Theta})$ can be derived by finding the extreme value of (38), as shown at the bottom of the page.

V. Parameter Update Equations

Having music applications in mind, let us assume here that all pitch contours are zero-order polynomials: $\mu_k(t) = \mu_{k0}$ and each partial stream in the HTC source model has the same power envelope (only a single power envelope function is assumed in the HTC source model): $U_{k,n}(t) = U_k(t)$, for the purpose of reducing the dimensionality of the features to extract. From (23), logarithmic subsource model $\log S_{k,n,y}(x,t;\boldsymbol{\Theta})$ is given by

$$\begin{aligned}\log S_{k,n,y}(x,t;\boldsymbol{\Theta}) = \log \frac{w_k v_{k,n} u_{k,y}}{2\pi\sigma_k\phi_k} \\ - \frac{(x - \mu_{k0} - \log n)^2}{2\sigma_k^2} \\ - \frac{(t - \tau_k - y\phi_k)^2}{2\phi_k^2}.\end{aligned} \tag{39}$$

$$\begin{aligned}\mathcal{I}(\boldsymbol{\Theta}) \triangleq \sum_k \Bigg(\sum_n \sum_y \iint_D m_k(x,t) m_{k,n,y}(x,t) W(x,t) \log S_{k,n,y}(x,t;\boldsymbol{\Theta}) \mathrm{d}x\mathrm{d}t + \sum_n d_v \bar{v}_n \log v_{k,n} + \sum_n \sum_y d_u \bar{u}_y \log u_{k,n,y} \\ - \gamma_v^{(k)} \Big(\sum_n v_{k,n} - 1 \Big) - \sum_n \gamma_u^{(k,n)} \Big(\sum_y u_{k,n,y} - 1 \Big) \Bigg) - \gamma_w \Big(\sum_k w_k - 1 \Big)\end{aligned} \tag{38}$$

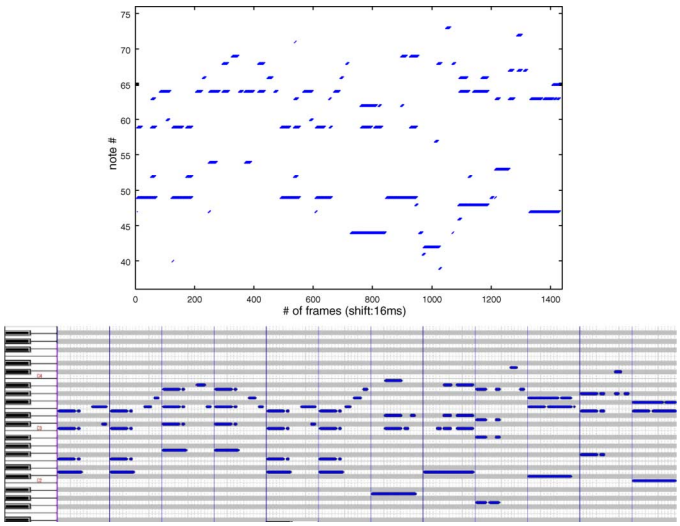


Fig. 6. Estimates of $\mu_{k0}, \tau_k, Y\phi_k$ (top), and the reference MIDI data displayed in piano-roll form (bottom).

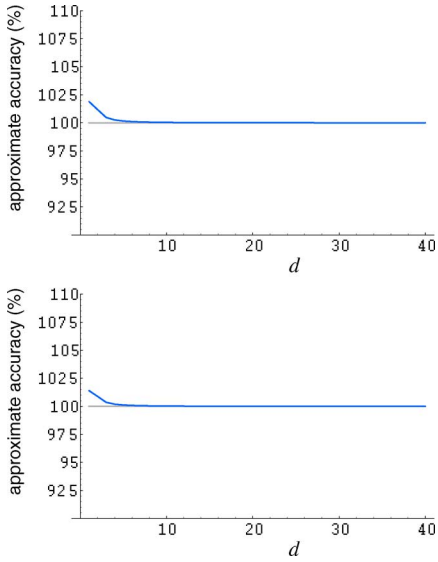


Fig. 7. Accuracy $\widehat{\mu}/\omega_0 \times 100$ (percent) of the result for each time spread parameter d ranging from 1 to 40 when ω_0 is set to $2\pi \times 100$ (top) and $2\pi \times 800$ (bottom) rad/s.

Setting to zero the partial derivative of (38), the update equation of each parameter at M-step of the ith iteration is derived as follows (see appendix for its derivation):

$$w_k^{(i)} = \sum_{n,y} \iint_D \ell_{k,n,y}^{(i)}(x,t)\mathrm{d}x\mathrm{d}t \tag{40}$$

$$\mu_{k0}^{(i)} = \frac{1}{w_k^{(i)}} \sum_{n,y} \iint_D (x - \log n)\ell_{k,n,y}^{(i)}(x,t)\mathrm{d}x\mathrm{d}t \tag{41}$$

$$\tau_k^{(i)} = \frac{1}{w_k^{(i)}} \sum_{n,y} \iint_D (t - y\phi_k^{(i-1)})\ell_{k,n,y}^{(i)}(x,t)\mathrm{d}x\mathrm{d}t \tag{42}$$

$$v_{k,n}^{(i)} = \frac{1}{d_v + w_k^{(i)}} \left(d_v \bar{v}_n + \sum_y \iint_D \ell_{k,n,y}^{(i)}(x,t)\mathrm{d}x\mathrm{d}t \right) \tag{43}$$

$$u_{k,y}^{(i)} = \frac{1}{d_u + w_k^{(i)}} \left(d_u \bar{u}_y + \sum_n \iint_D \ell_{k,n,y}^{(i)}(x,t)\mathrm{d}x\mathrm{d}t \right) \tag{44}$$

$$\phi_k^{(i)} = \frac{1}{2w_k^{(i)}} \left(\left(a_k^{(i)2} + 4b_k^{(i)} w_k^{(i)} \right)^{1/2} - a_k^{(i)} \right) \tag{45}$$

$$\begin{cases} a_k^{(i)} \triangleq \sum_{n,y} \iint_D y(t - \tau_k^{(i)})\ell_{k,n,y}^{(i)}(x,t)\mathrm{d}x\mathrm{d}t \\ b_k^{(i)} \triangleq \sum_{n,y} \iint_D (t - \tau_k^{(i)})^2 \ell_{k,n,y}^{(i)}(x,t)\mathrm{d}x\mathrm{d}t \end{cases}$$

$$\sigma_k^{(i)} = \left(\frac{1}{w_k^{(i)}} \sum_{n,y} \iint_D (x - \mu_{k0}^{(i)} - \log n)^2 \ell_{k,n,y}^{(i)}(x,t)\mathrm{d}x\mathrm{d}t \right)^{1/2} \tag{46}$$

where ℓ denotes the subcluster densities

$$\ell_{k,n,y}^{(i)}(x,t) = m_k^{(i)}(x,t) m_{k,n,y}^{(i)}(x,t) W(x,t).$$

The superscript (i) refers to the iteration cycle. Note that the update equations of the higher order coefficients $\mu_{k1}, \mu_{k2}, \ldots$, of the pitch trajectory model can all be derived as well as μ_{k0}, if necessary.

Note, however, that the update (40)–(46) do not ensure the maximization in the M-step but guarantee the increase of $\widetilde{J}^+ + \log P(\boldsymbol{\Theta})$. This way of performing the updates is referred to as the coordinate descent method [12] and the corresponding optimization procedure is called the expectation-constrained maximization (ECM) algorithm [13].

VI. Relation With PreFEst [10] and Our Previous Work [11]

A cutting plane of the HTC source model at a particular time gives a similar representation to the harmonic structure model introduced by Goto in [10] and followed independently by our previous work described in [11]. Goto used a tone model, a harmonic structure model represented by a Gaussian mixture, and tried to model an observed short-time power spectrum by the mixture of a large enough number of the tone models densely spaced by a fixed interval. He then used the EM algorithm frame by frame for estimating the MAP mixture weights of the tone models to measure pitch likelihoods from an observed power spectrum. In our previous work described in [11], on the other hand, we started with a formulation based on the clustering principle using a harmonic structured Gaussian mixture cluster model in order to try to directly estimate each mean parameter, the pitch estimate itself. The source spectrogram model introduced in the HTC method designs the overall smooth temporal evolution of the spectral structure model introduced by us and Goto.

VII. Experimental Evaluation

A. Conditions

To verify the performance of the HTC method, we evaluated pitch estimation accuracies with a set of real performed music signals excerpted from the RWC music database [14]. The first 23 s of each music signal data were used for the evaluation. We implemented a GUI editor to create a ground truth data set of pitch sequences (a screenshot of the GUI editor can be seen in Fig. 9). Since the RWC database also includes a supplement MIDI file associated with each real-performed music signal data, we created each ground truth data set with the GUI

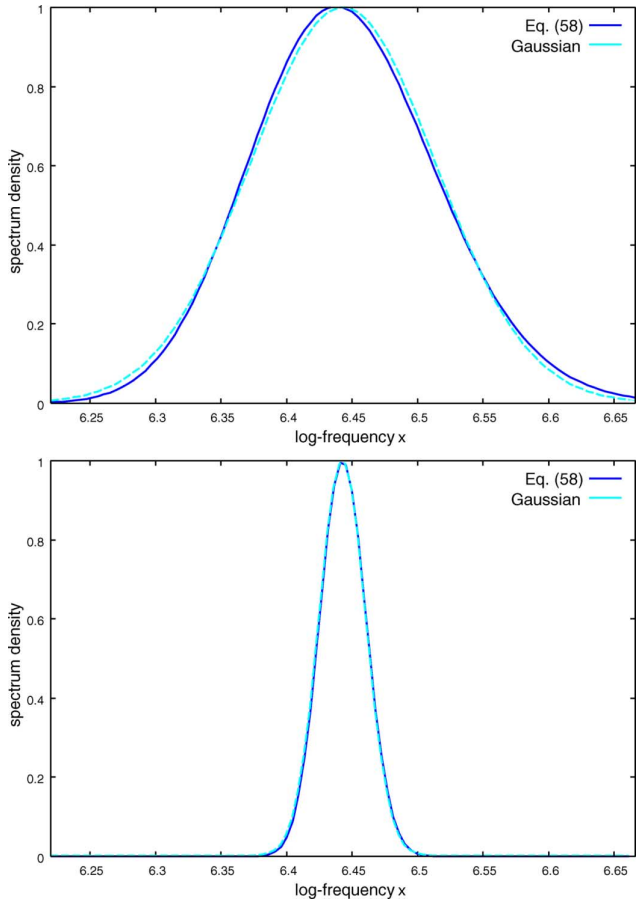


Fig. 8. Graphical representations of (62) with $d = 10$ (top) and $d = 40$ (bottom) when ω_0 is set to $2\pi \times 100$ rad/s together with the fitted Gaussians.

Fig. 9. Screenshot of the GUI editor we implemented to create the ground truth data set of note pitches, onsets, and durations. The note events of the supplement MIDI data included in the RWC database, which are not temporally aligned with the corresponding real performed signal data, are displayed as rectangular objects over the spectrogram of the real performed signal. We are then able to edit the rectangular objects to align carefully the onset and offset times according to the background spectrogram.

by adjusting the onset and offset times of each MIDI file with a spectrogram image for a background. The list of the experimental data sets and the time average of the number of concurrent sources, the total number of the frames for all pitches divided by the number of the frames, are shown in Table II.

The Power spectrum time series was analyzed by the wavelet transform (constant Q analysis) using Gabor-wavelet basis functions with a time resolution of 16 ms for the lowest frequency subband on an input signal digitalized at a 16-kHz sampling rate. To speed up the computation time, we set the time resolution across all the subbands equally to 16 ms. The lower bound of the frequency range and the frequency resolution were 60 Hz and 12 cents (where 12 cents amount to one octave), respectively. The initial parameters of $(\mu_{k0}, \tau_k | k = 1, \ldots, K)$ for the HTC source models were automatically determined by picking the 60 largest peaks in the observed spectrogram of 400 consecutive frames (6.4 s). After the parameters converged, the source model, whose energy per unit time given by $w_k / Y \phi_k$ was smaller than a threshold, was considered to be silent. The experimental conditions are shown in detail in Table III.

We chose[2] "PreFEst" [10] for comparison, as it is one of the most frequently cited works dedicated to multipitch analysis. Since PreFEst extracts only the most dominant pitch trajectory and does not include a specific procedure of estimating the number of sources, we included intensity thresholding as well for the pitch candidate truncation.

As the HTC method generates pitch, onset time and offset time with continuous values, we quantize them to the closest note and the closest frame number in order to match with the format of the reference. Using the hand-labeled ground truth data as references, pitch accuracies were computed by

$$\frac{X - D - I - S}{X} \times 100(\%)$$

where

- X number of the total frames of the voiced part;
- D number of deletion errors;
- I number of insertion errors;
- S number of substitution errors.

B. Results

A typical example of the pitch, onset, and offset estimates on a particular test data is shown in Fig. 6 together with the hand-labeled ground truth data. The optimized model and the observed power spectrum time series are shown with 3-D and grayscale displays in Fig. 5.

To validate the performance of the proposed method, we compared the highest accuracy of the HTC method with that of the PreFEst among all the thresholds that were tested, which also shows the limit of the potential capability. The highest accuracies of PreFEst and HTC among all the thresholds we tested are shown in Table IV together with the number of insertion, deletion, and substitution errors, respectively. Comparing these accuracies between PreFEst and HTC, HTC outperforms PreFEst for most of the data, which verifies its potential.

The workstation used to perform the experiments had a Pentium IV processor with 3.2-GHz clock speed and 2-GB memory. With our implementation with the conditions listed in Table III,

[2]Note that we implemented for the evaluation only the module called "PreFEst-core," a frame-wise pitch likelihood estimation, and not included the one called "PreFEst-back-end," a multiagent-based pitch tracking algorithm. Refer to [10] for their details.

TABLE II
LIST OF THE EXPERIMENTAL DATA EXCERPTED FROM RWC MUSIC DATABASE.[14]

Symbol	Title (Genre)	Catalog number	Composer/Player	Instruments	Ave. # of sources
data(1)	Crescent Serenade (Jazz)	RWC-MDB-J-2001 No. 9	S. Yamamoto	Guitar	2.13
data(2)	For Two (Jazz)	RWC-MDB-J-2001 No. 7	H. Chubachi	Guitar	2.67
data(3)	Jive (Jazz)	RWC-MDB-J-2001 No. 1	M. Nakamura	Piano	1.86
data(4)	Lounge Away (Jazz)	RWC-MDB-J-2001 No. 8	S. Yamamoto	Guitar	4.04
data(5)	For Two (Jazz)	RWC-MDB-J-2001 No. 2	M. Nakamura	Piano	2.34
data(6)	Jive (Jazz)	RWC-MDB-J-2001 No. 6	H. Chubachi	Guitar	1.78
data(7)	Three Gimnopedies no. 1 (Classic)	RWC-MDB-C-2001 No. 35	E. Satie	Piano	2.96
data(8)	Nocturne no.2, op.9-2(Classic)	RWC-MDB-C-2001 No. 30	F. F. Chopin	Piano	1.55

TABLE III
EXPERIMENTAL CONDITIONS

frequency analysis	Sampling rate	16 kHz
	frame shift	16 ms
	frequency resolution	12.0 cent
	frequency range	60–3000 Hz
HTC	# of HTC source models: K	60
	# of partials: N	6
	# of kernels in $U_k(t)$: Y	10
	$\bar{v}_n$	$0.6547 \times n^{-2}$
	$\bar{u}_y$	$0.2096 \times e^{-0.2y}$
	d_v, d_u	0.04
	time range of a spectrogram segment	400 frames (6.4 s)
	# of the segments	4 (total time: 25.6 s)
PreFEst [10]	pitch resolution	20 cent
	# of partials	8
	# of tone models	200
	standard deviation of Gaussian	3.0
	$\bar{r}_n$	$0.6547 \times n^{-2}$
	d (prior contribution factor)	3.0

the computational time for analyzing an acoustic signal of 25.6-s length was about 2 min. In most cases, the parameters of the HTC source models converged within less than 100 iteration cycles.

We also compared the HTC performances with different conditions: the time range of an analyzing spectrogram segment of 100, 200, and 400 frames, and the number of the HTC source models of 15, 30, and 60, respectively. Comparative results are shown in Table V. From the results, one can see that the larger the time range of a spectrogram segment, the higher the accuracies. This shows that the domain of definition of t should be as large as possible for a higher performance of the HTC.

VIII. DISCUSSION OF THE POTENTIAL APPLICATIONS

A. Sound Source Segregation

It should be emphasized that the HTC method can also be used to extract a portion of the power spectrum time series associated with a single source by

$$\frac{q_k(x,t;\widehat{\boldsymbol{\Theta}})}{\sum_k q_k(x,t;\widehat{\boldsymbol{\Theta}})} W(x,t) \tag{47}$$

where $\widehat{\boldsymbol{\Theta}}$ is the optimized model parameter vector. Using a decoding technique (such as the inverse wavelet transform, phase vocoder, etc.) to reconstruct acoustic signals from the power spectrum time series, the HTC also enables the separation of sources.

B. Audio Coding

Since $\sum_k q_k(x,t;\widehat{\boldsymbol{\Theta}})$ must be a good approximation to the observed power spectrum time series where $\widehat{\boldsymbol{\Theta}}$ is the optimal model parameter vector, the residual power spectrum

$$W(x,t) - \sum_k q_k(x,t;\widehat{\boldsymbol{\Theta}}) \tag{48}$$

may be effectively compressed by the well-known Huffman encoding. As the original power spectrum can be restored with the residual and the optimized parameter vector, our method is thus also expected to be used as an effective audio coding application if acoustic signals can be intelligibly reconstructed from power spectra. However, its effectiveness over traditional lossless and lossy coding techniques will need to be evaluated in future investigations.

IX. CONCLUSION

We established a new framework for multipitch analysis based upon two-dimensional geometric modeling and estimation of the distinct spectral streams localized in the time-frequency space, the acoustic scene, and investigated through experiments its effectiveness over the conventional method, PreFEst.

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 of pitch contour, and the inharmonicity factor parameter and a further investigation on its application.

APPENDIX I
JUSTIFICATION FOR THE ASSUMPTION OF THE FREQUENCY SPREAD BEING A GAUSSIAN DISTRIBUTION

Denoting by $f(t)$ and $F(\omega)$ a given signal and its Fourier transform, we define the Fourier transform pair as follows:

$$F(\omega) = \mathcal{F}[f(t)] \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-j\omega t} \mathrm{d}t \tag{49}$$

$$f(t) = \mathcal{F}^{-1}[F(\omega)] \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega) e^{j\omega t} \mathrm{d}\omega. \tag{50}$$

Denoting by $\varphi_{a,b}(t)$ the wavelet basis function

$$\varphi_{a,b}(t) = \frac{1}{\sqrt{a}} \varphi\left(\frac{t-b}{a}\right) \tag{51}$$

where $\varphi(t)$ is the analyzing wavelet. a and b refer to the scale and shift parameters, which have the dimensions of period and

TABLE IV
ACCURACIES OF THE PREFEST [10] AND THE HTC

		conventional 'PreFEst'[10]				*proposed 'HTC'*			
	X	Accuracy (%)	I	D	S	Accuracy (%)	I	D	S
data(1)	3063	74.2	383	327	81	**81.2**	210	312	55
data(2)	3828	71.8	455	397	228	**77.9**	241	397	208
data(3)	2671	55.9	553	500	126	**64.2**	313	524	120
data(4)	5798	**76.2**	476	650	254	75.2	361	769	310
data(5)	3366	**62.3**	565	515	190	62.2	465	627	178
data(6)	2563	48.8	531	597	185	**63.8**	304	476	147
data(7)	4244	53.6	801	830	337	**63.2**	427	734	403
data(8)	2227	57.6	367	482	96	**70.9**	278	291	79

TABLE V
COMPARISON OF THE HTC PERFORMANCES WITH DIFFERENT RANGES OF A SPECTROGRAM SEGMENT AND THE NUMBER OF SOURCE MODELS

		Time range: 100 frames, K: 15				*Time range: 200 frames, K: 30*				*Time range: 400 frames, K: 60*			
	X	Accuracy (%)	I	D	S	Accuracy (%)	I	D	S	Accuracy (%)	I	D	S
data(1)	3063	68.5	130	677	159	79.4	188	368	76	**81.2**	210	312	55
data(2)	3828	75.1	142	720	93	74.2	218	538	233	**77.9**	241	397	208
data(3)	2671	58.7	271	671	160	61.8	332	549	139	**64.2**	313	524	120
data(4)	5798	60.7	175	1863	243	66.6	232	1376	327	**75.2**	361	769	310
data(5)	3366	55.3	427	926	153	59.6	385	774	201	**62.2**	465	627	178
data(6)	2563	57.7	229	617	239	61.2	270	519	206	**63.8**	304	476	147
data(7)	4244	54.4	309	1226	400	**63.5**	470	619	461	63.2	427	734	403
data(8)	2227	58.8	234	598	85	68.2	315	325	69	**70.9**	278	291	79

time, respectively. Based on the above definitions and the generalized Parseval's theorem, the continuous wavelet transform is defined in the frequency domain as an inner product between $F(\omega)$ and $\Psi_{a,b}(\omega)$, the Fourier transform of $\varphi_{a,b}(t)$

$$W(a,b) \triangleq \Big\langle f(t), \varphi_{a,b}(t) \Big\rangle = \Big\langle F(\omega), \Psi_{a,b}(\omega) \Big\rangle. \tag{52}$$

Denoting by $\Psi(\omega)$ the Fourier transform of $\varphi(t)$ such that

$$\Psi_{a,b}(\omega) = \sqrt{a}\Psi(a\omega)e^{j\omega b} \tag{53}$$

then the wavelet transform amounts to an inverse Fourier transform of the subband-filtered spectrum $\sqrt{a}\Psi^*(a\omega)F(\omega)$

$$W(a,b) = \int_{-\infty}^{\infty} \sqrt{a}\Psi^*(a\omega)F(\omega)e^{j\omega b}\mathrm{d}\omega. \tag{54}$$

Let us now consider the very simple case where $f(t)$ is a sinusoidal wave and see how the frequency spread at a certain time slice of the wavelet scalogram is given by. Letting $f(t)$ be an analytic signal with an angular frequency of ω_0 and with an amplitude of $\xi > 0$

$$f(t) = \xi e^{j\omega_0 t} \tag{55}$$

its Fourier transform is then given by

$$F(\omega) = \xi\delta(\omega - \omega_0) \tag{56}$$

where $\delta(\cdot)$ denotes the Dirac delta function. Substituting (56) into (54), one immediately obtains

$$W(a,b) = \xi\sqrt{a}\Psi^*(a\omega_0)e^{j\omega_0 b}. \tag{57}$$

Its power is thus given by

$$|W(a,b)|^2 = \xi^2 a|\Psi^*(a\omega_0)|^2 \tag{58}$$

from which we obviously see that $|W(a,b)|^2$ does not depend on b and is thus uniform over time.

Now, if we choose for $\varphi(t)$ a Gabor function of frequency 1 rad/s with time spread $d > 0$

$$\varphi(t) = e^{-(t^2/2d^2)+jt} \tag{59}$$

$\Psi(\omega)$ is then a Gaussian function centered at 1 rad/s

$$\Psi(\omega) = \Psi^*(\omega) = Ce^{-(d^2/2)(\omega-1)^2}. \tag{60}$$

$\sqrt{a}\Psi^*(a\omega)$ is thus a Gaussian subband filter with center frequency of $1/a$ rad/s. Substituting (60) into (58), the power density in the period domain is given explicitly as

$$|W(a,b)|^2 = C^2\xi^2 a e^{-d^2(a\omega_0-1)^2}. \tag{61}$$

The density $g(x)$ in the log-frequency domain x such that $x = \log(1/a)$ at a particular time slice of $|W(e^{-x},b)|^2$ is thus given as

$$g(x) = C^2\xi^2 \exp\Big(-x - d^2\big(\omega_0 e^{-x} - 1\big)^2\Big) \tag{62}$$

which is obviously not a Gaussian distribution function.

In order to justify the assumption made in the HTC source model (Section III) that the frequency spread of the wavelet power spectra is close to a Gaussian distribution, we will now investigate through a numerical computation how much the mean parameter of the Gaussian distribution that best approximates the above function with regard to the KL-divergence criterion deviates from the true frequency ω_0. Denoting by $q(x)$ the Gaussian distribution model of the mean parameter μ

$$q(x) = \frac{D}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2} \tag{63}$$

the KL divergence between g and q is defined as

$$\mathcal{J} \triangleq \int_{-\infty}^{\infty} g(x) \log \frac{g(x)}{q(x)} \mathrm{d}x \tag{64}$$

with

$$\int_{-\infty}^{\infty} g(x) \mathrm{d}x = \int_{-\infty}^{\infty} q(x) \mathrm{d}x. \tag{65}$$

The optimal μ, that is supposed to be here the estimate of ω_0, can thus be computed by finding the extreme value of $\mathcal{J}$. Setting to zero the partial derivative of $\mathcal{J}$ with respect to μ, which is given as

$$\frac{\partial \mathcal{J}}{\partial \mu} = - \int_{-\infty}^{\infty} g(x) \frac{x-\mu}{\sigma^2} \mathrm{d}x \tag{66}$$

one obtains

$$\widehat{\mu} = \frac{\int_{-\infty}^{\infty} x g(x) \mathrm{d}x}{\int_{-\infty}^{\infty} g(x) \mathrm{d}x}. \tag{67}$$

Since it is impossible to obtain analytically both the integrals in the numerator and denominator in (67), we give numerical results for $\widehat{\mu}$ when ω_0 is set to $2\pi \times 100$ and $2\pi \times 800$ rad/s, respectively. Fig. 7 shows the accuracy $\widehat{\mu}/\omega_0 \times 100$ (%) of the result for each time spread parameter d ranging from 1 to 40. When $d = 40$, the value we chose in the experiment described in Section VII, the accuracy for both $\widehat{\mu}/\omega_0 \times 100$ (%) with $\omega_0 = 2\pi \times 100$ and $\omega_0 = 2\pi \times 800$ were around 100.002 (%). This numerical result shows that the assumption of the frequency spread being a Gaussian distribution may not necessarily affect critically the result of the pitch estimations. Graphical representations of (62) with $d = 10$ and $d = 40$ can be seen in Fig. 8, from which one is able to see how each of them is close to a Gaussian distribution.

Appendix II
Derivation of the Parameter Update Equations

We will show in this section how the parameter update (40)–(46) were derived.

Energy: w_k: From (4), one immediately obtains

$$w_k^{(i)} = \sum_{n,y} \iint_D \ell_{k,n,y}(x,t) \mathrm{d}x\mathrm{d}t. \tag{68}$$

Constant Term of the Pitch Contour: μ_{k0}: Setting to zero the partial derivative of $\mathcal{I}(\boldsymbol{\Theta})$ w.r.t. μ_{k0}

$$\frac{\partial \mathcal{I}(\boldsymbol{\Theta})}{\partial \mu_{k0}} = \sum_{n,y} \iint_D \frac{x - \mu_{k0} - \log n}{\sigma_k^2} \ell_{k,n,y}(x,t) \mathrm{d}x\mathrm{d}t = 0 \tag{69}$$

one obtains

$$\mu_{k0}^{(i)} = \frac{\sum\limits_{n,y} \iint_D (x - \log n) \ell_{k,n,y}(x,t) \mathrm{d}x\mathrm{d}t}{\sum\limits_{n,y} \iint_D \ell_{k,n,y}(x,t) \mathrm{d}x\mathrm{d}t}. \tag{70}$$

As we find that the denominator of the above is equal to (68), one finally obtains

$$\mu_{k0}^{(i)} = \frac{1}{w_k^{(i)}} \sum_{n,y} \iint_D (x - \log n) \ell_{k,n,y}(x,t) \mathrm{d}x\mathrm{d}t. \tag{71}$$

Onset Time: τ_k: Setting to zero the partial derivative of $\mathcal{I}(\boldsymbol{\Theta})$ w.r.t. τ_k

$$\frac{\partial \mathcal{I}(\boldsymbol{\Theta})}{\partial \tau_k} = \sum_{n,y} \iint_D \frac{t - \tau_k - y\phi_k}{\phi_k^2} \ell_{k,n,y}(x,t) \mathrm{d}x\mathrm{d}t = 0 \tag{72}$$

one obtains

$$\tau_k^{(i)} = \frac{1}{w_k^{(i)}} \sum_{n,y} \iint_D (t - y\phi_k^{(i-1)}) \ell_{k,n,y}(x,t) \mathrm{d}x\mathrm{d}t. \tag{73}$$

Relative Energy of the Harmonics: $v_{k,n}$: Setting to zero the partial derivative of $\mathcal{I}(\boldsymbol{\Theta})$ w.r.t. $v_{k,n}$

$$\frac{\partial \mathcal{I}(\boldsymbol{\Theta})}{\partial v_{k,n}} = \frac{1}{v_{k,n}} \sum_{y} \iint_D \ell_{k,n,y}(x,t) \mathrm{d}x\mathrm{d}t + \frac{d_v \bar{v}_n}{v_{k,n}} - \gamma_v^{(k)} = 0 \tag{74}$$

one obtains

$$v_{k,n}^{(i)} = \frac{1}{\gamma_v^{(k)}} \left(\sum_{y} \iint_D \ell_{k,n,y}(x,t) \mathrm{d}x\mathrm{d}t + d_v \bar{v}_{k,n} \right). \tag{75}$$

From (75) and (18), the Lagrange multiplier $\gamma_v^{(k)}$ is given explicitly as

$$\gamma_v^{(k)} = \sum_{n,y} \iint_D \ell_{k,n,y}(x,t) \mathrm{d}x\mathrm{d}t + d_v \tag{76}$$

which finally gives us the following:

$$v_{k,n}^{(i)} = \frac{1}{d_v + w_k^{(i)}} \left(d_v \bar{v}_{k,n} + \sum_{y} \iint_D \ell_{k,n,y}(x,t) \mathrm{d}x\mathrm{d}t \right). \tag{77}$$

Coefficients of the Power Envelope Function: $u_{k,y}$: Setting to zero the partial derivative of $\mathcal{I}(\boldsymbol{\Theta})$ w.r.t. $u_{k,y}$

$$\frac{\partial \mathcal{I}(\boldsymbol{\Theta})}{\partial u_{k,y}} = \frac{1}{u_{k,y}} \sum_{n} \iint_D \ell_{k,n,y}(x,t) \mathrm{d}x\mathrm{d}t + \frac{d_u \bar{u}_y}{u_{k,y}} - \gamma_u^{(k)} = 0, \tag{78}$$

one obtains

$$u_{k,y}^{(i)} = \frac{1}{\gamma_u^{(k)}} \left(\sum_{n} \iint_D \ell_{k,n,y}(x,t) \mathrm{d}x\mathrm{d}t + d_u \bar{u}_{k,y} \right). \tag{79}$$

From (79) and (21), the Lagrange multiplier $\gamma_u^{(k)}$ is given explicitly as

$$\gamma_u^{(k)} = \sum_{n,y} \iint_D \ell_{k,n,y}(x,t) \mathrm{d}x\mathrm{d}t + d_u \tag{80}$$

which finally gives us the following:

$$u_{k,y}^{(i)} = \frac{1}{d_u + w_k^{(i)}} \left(d_u \bar{u}_{k,y} + \sum_n \iint_D \ell_{k,n,y}(x,t) \mathrm{d}x\mathrm{d}t \right). \tag{81}$$

Duration: $Y\phi_k$: Setting to zero the partial derivative of $\mathcal{I}(\boldsymbol{\Theta})$ w.r.t. ϕ_k

$$\frac{\partial \mathcal{I}(\boldsymbol{\Theta})}{\partial \phi_k} = \sum_{n,y} \iint_D \frac{(t-\tau_k)(t-\tau_k-y\phi_k)-\phi_k^2}{\phi_k^3} \ell_{k,n,y}(x,t)\mathrm{d}x\mathrm{d}t = 0 \tag{82}$$

can be rewritten in the form (a quadratic equation)

$$w_k^{(i)}\phi_k^2 + a\phi_k - b = 0 \tag{83}$$

where

$$\begin{cases} a = \sum\limits_{n,y} \iint_D y(t-\tau_k)\ell_{k,n,y}(x,t)\mathrm{d}x\mathrm{d}t \\ b = \sum\limits_{n,y} \iint_D (t-\tau_k)^2 \ell_{k,n,y}(x,t)\mathrm{d}x\mathrm{d}t \end{cases}$$

with $\phi_k^{(i)} > 0$, from which one finally obtains

$$\phi_k^{(i)} = \frac{-a + \left(a^2 + 4bw_k^{(i)}\right)^{1/2}}{2w_k^{(i)}}. \tag{84}$$

Frequency Spread of Partial Distributions: σ_k: Setting to zero the partial derivative of $\mathcal{I}(\boldsymbol{\Theta})$ w.r.t. σ_k

$$\frac{\partial \mathcal{I}(\boldsymbol{\Theta})}{\partial \sigma_k} = \sum_{n,y} \iint_D \frac{(x-\mu_{k0}-\log n)^2 - \sigma_k^2}{\sigma_k^3} \ell_{k,n,y}(x,t)\mathrm{d}x\mathrm{d}t = 0 \tag{85}$$

one obtains

$$\sigma_k^{(i)} = \left(\frac{\sum\limits_{n,y} \iint_D (x-\mu_{k0}^{(i)}-\log n)^2 \ell_{k,n,y}(x,t)\mathrm{d}x\mathrm{d}t}{w_k^{(i)}} \right)^{1/2}. \tag{86}$$

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their valuable comments, Dr. M. Goto, J. Le Roux, and Dr. N. Ono for their fruitful discussions on this work, and Dr. M. Schuster for his help with the English language.

REFERENCES

[1] 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* , 1995, vol. 1, pp. 158–164.

[2] A. Klapuri, T. Virtanen, and J. Holm, "Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals," in *Proc. COST-G6 Conf. Digital Audio Effects*, 2000, pp. 233–236.

[3] K. Nishi and S. Ando, "Optimum harmonics tracking filter for auditory scene analysis," in *Proc. IEEE ICASSP '96*, 1996, vol. 1, pp. 573–576.

[4] 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," (in Japanese) *Trans. IEICE*, vol. J83-D-II, no. 2, pp. 468–477, 2000.

[5] T. Nakatani, "Computational auditory scene analysis based on residue-driven architecture and its application to mixed speech recognition," Ph.D. dissertation, Kyoto Univ., Kyoto, Japan, 2002.

[6] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 229–241, May 2003.

[7] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 4, pp. 477–489, Apr. 1988.

[8] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation," in *Proc. IEEE ICASSP*, 1993, vol. 2, pp. 728–731.

[9] S. Godsill and M. Davy, "Baysian harmonic models for musical pitch estimation and analysis," in *Proc. IEEE ICASSP*, 2002, vol. 2, pp. 1769–1772.

[10] M. Goto, "A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *ISCA J.*, vol. 43, no. 4, pp. 311–329, 2004.

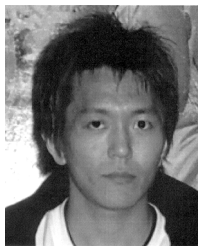
[11] H. Kameoka, T. Nishimoto, and S. Sagayama, "Separation of harmonic structures based on tied Gaussian mixture model and information criterion for concurrent sounds," in *Proc. IEEE ICASSP*, 2004, vol. 4, pp. 297–300.

[12] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1969.

[13] X. L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.

[14] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music database," in *Proc. ISMIR*, 2002, pp. 287–288.

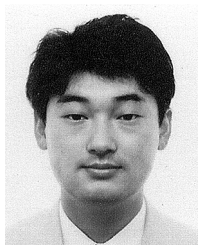
[15] H. Kameoka, T. Nishimoto, and S. Sagayama, "Audio stream segregation of multi-pitch music signals based on time-space clustering using Gaussian kernel 2-dimensional model," in *Proc. IEEE ICASSP*, 2005, vol. 3, pp. 5–8.



Hirokazu Kameoka (S'05) received the B.E. and M.E. degrees from University of Tokyo, Tokyo, Japan, in 2002 and 2004, respectively. He is currently pursuing the Ph.D. degree at the Graduate School of Information Science and Technology, University of Tokyo.

His research interests include acoustic signal processing, speech processing, and music processing.

Mr. Kameoka is a student member of the Institute of Electronics, Information and Communication Engineers (IEICE), IPSJ, Acoustical Society of Japan (ASJ), and International Speech Communication Association (ISCA). He was awarded the Yamashita Memorial Research Award from the Information Processing Society of Japan (IPSJ), Best Student Paper Award Finalist at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), and the 20th Telecom System Technology Student Award from the Telecomunications Advancement Foundation (TAF), all in 2005.



Takuya Nishimoto received the B.E. and M.E. degrees from Waseda University, Tokyo, Japan, in 1993 and 1995, respectively.

He is a Research Associate at the Graduate School of Information Science and Technology, University of Tokyo. His research interests include spoken dialogue systems and human–machine interfaces.

Mr. Nishimoto is a member of the Institute of Electronics, Information, and Communication Engineers (IEICE), Japan, Information Processing Society of Japan (IPSJ), Acoustical Society of Japan (ASJ), Japanese Society for Artificial Intelligence (JSAI), and Human Interface Society (HIS).

Shigeki Sagayama (M'82) was born in Hyogo, Japan, in 1948. He received the B.E., M.E., and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 1972, 1974, and 1998, respectively, all in mathematical engineering and information physics.

He joined Nippon Telegraph and Telephone Public Corporation (currently, NTT) in 1974 and started his career in speech analysis, synthesis, and recognition at NTT Laboratories, Musashino, Japan. From 1990 to 1993, he was Head of Speech Processing Department, ATR Interpreting Telephony Laboratories, Kyoto, Japan, pursuing an automatic speech translation project. From 1993 to 1998, he was responsible for speech recognition, synthesis, and dialog systems at NTT Human Interface Laboratories, Yokosuka, Japan. In 1998, he became a Professor of the Graduate School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan. In 2000, he was appointed Professor of the Graduate School of Information Science and Technology (formerly Graduate School of Engineering), University of Tokyo. His major research interests include processing and recognition of speech, music, acoustic signals, hand writing, and images. He was the leader of anthropomorphic spoken dialog agent project (Galatea Project) from 2000 to 2003.

Prof. Sagayama is a member of the Acoustical Society of Japan (ASJ), IEICEJ, and IPSJ. He received the National Invention Award from the Institute of Invention of Japan in 1991, the Chief Official's Award for Research Achievement from the Science and Technology Agency of Japan in 1996, and other academic awards including Paper Awards from the Institute of Electronics, Information, and Communications Engineers (IEICEJ) Japan, in 1996 and from the Information Processing Society of Japan (IPSJ) in 1995.