INFINITE LATENT HARMONIC ALLOCATION: A NONPARAMETRIC BAYESIAN APPROACH TO MULTIPITCH ANALYSIS

Kazuyoshi Yoshii **Masataka Goto**
National Institute of Advanced Industrial Science and Technology (AIST), Japan
{k.yoshii, m.goto}@aist.go.jp

ABSTRACT

This paper presents a statistical method called *Infinite Latent Harmonic Allocation* (iLHA) for detecting multiple fundamental frequencies in polyphonic audio signals. Conventional methods face a crucial problem known as model selection because they assume that the observed spectra are superpositions of a *certain fixed* number of bases (sound sources and/or finer parts). iLHA avoids this problem by assuming that the observed spectra are superpositions of a *stochastically-distributed unbounded* (theoretically infinite) number of bases. Such uncertainty can be treated in a principled way by leveraging the state-of-the-art paradigm of machine-learning called Bayesian nonparametrics. To represent a set of time-sliced spectral strips, we formulated nested infinite Gaussian mixture models (GMMs) based on hierarchical and generalized Dirichlet processes. Each strip is allowed to contain an unbounded number of sound sources (GMMs), each of which is allowed to contain an unbounded number of harmonic partials (Gaussians). To train the nested infinite GMMs efficiently, we used a modern inference technique called collapsed variational Bayes (CVB). Our experiments using audio recordings of real piano and guitar performances showed that fully automated iLHA based on noninformative priors performed as well as optimally tuned conventional methods.

1. INTRODUCTION

Multipitch analysis of polyphonic audio signals [1–11] is one of the most important issues because it is the basis of many applications such as music transcription, chord recognition, and musical instrument recognition. We focus on principled methods based on machine learning, which have recently yielded promising results. Some researchers, for example, have proposed generative probabilistic models that explain how multiple spectral/signal bases (compositional units) are mixed to form polyphonic music [3–6]. The model parameters can be trained by means of statistical inference. Others have used nonnegative matrix factorization (NMF) to decompose polyphonic spectra into individual spectral bases [7–11]. NMF can be interpreted from the viewpoint of statistical inference [10–12].

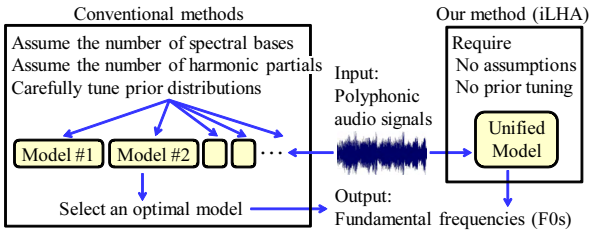

Figure 1. Methodological advantage of our method.

A crucial problem in these methods, known as model selection, is that they perform best only if an appropriate model complexity (the number of bases) is specified *in advance*. One might think that the optimal number of bases must be equal to the number of sound sources, but it is not clear how many bases are most suited to represent a single source if the spectral shape varies through time. Although *uncertainty* is inherent in model selection, conventional methods assume that a *certain* complexity exists uniquely as an oracle. As shown in Figure 1, they require possible models to be examined separately and exhaustively and the optimal model selected *in retrospect*. Such a deterministic framework is not easy-to-use in practice although optimally tuned methods can achieve good performance.

To avoid model selection, we propose a novel statistical method called *Infinite Latent Harmonic Allocation* (iLHA) based on a modern paradigm of machine learning called Bayesian nonparametrics. Note that the term "nonparametric" means that we do not have to fix model complexity uniquely. We assume that an unbounded but finite number of bases stochastically appears in a limited amount of available data although an infinite number of bases theoretically exists in the universe. Uncertainty in model selection can be treated reasonably in a probabilistic framework.

iLHA can be derived by taking the infinite limit of conventional finite models [3, 4]. Conventionally, each spectral basis is often parameterized by means of a Gaussian mixture model (GMM) in which a fixed number of Gaussians corresponds to the spectral peaks of harmonic partials, and a time-sliced polyphonic spectral strip is modeled by mixing a fixed number of GMMs. Here, we consider both the number of bases and the number of partials to approach infinity, where most are regarded as unnecessary and automatically removed through statistical inference.

A fundamental and practically-important advantage of iLHA is that precise prior knowledge is not required. Conventional methods [3–5] heavily rely on prior distributions regarding the relative strengths of harmonic partials, which

have too much impact on performance, and forced us to tune priors and their weighting factors by hand according to the properties of target sound sources. iLHA, in contrast, can be fully automated by layering noninformative hyperpriors on influential priors in a hierarchical Bayesian manner. This is consistent with the fact that humans can adaptively distinguish individual notes of various instruments. One of major contributions of this study is to embody the fundamental Bayesian principle "Let the data speak for itself" in the context of multipitch analysis.

The rest of this paper is organized as follows: Section 2 describes statistical interpretation of polyphonic spectra. Section 3 discusses related work. Sections 4 and 5 explain finite models (LHA) and infinite models (iLHA). Section 6 reports our experiments. Section 7 concludes this paper.

2. STATISTICAL INTERPRETATION

We interpret polyphonic spectra as histograms of observed frequencies that independently occur. This interpretation basically follows conventional studies [3–5].

2.1 Assumptions

Suppose given polyphonic audio signals are generated from K bases, each of which consists of M harmonic partials located on a linear frequency scale at integral multiples of the fundamental frequency (F0). Note that each *basis* can be associated with multiple *sounds* of different temporal positions if these sounds are derived from the same pitch of the same instrument. We transform the audio signals into wavelet spectra. Let D be the number of frames. If a spectral strip at frame d $(1 \leq d \leq D)$ has amplitude a at frequency f, we assume that frequency f was observed a times in frame d. Assuming that amplitudes are additive, we can consider each observed frequency to be generated from one of M partials in one of K bases.

These notations are for the finite case. In Bayesian nonparametrics, we take the limit as K and M go to infinity.

2.2 Observed and Latent Variables

Let the total observed variables over all D frames be represented by $\boldsymbol{X} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_D\}$, where $\boldsymbol{X}_d$ is a set of observed frequencies $\boldsymbol{X}_d = \{\boldsymbol{x}_{d1}, \cdots, \boldsymbol{x}_{dN_d}\}$ in frame d. N_d is the number of frequency observations. That is, N_d is equal to the sum of spectral amplitudes over all frequency bins in frame d. $\boldsymbol{x}_{dn}$ $(1 \leq n \leq N_d)$ is an one-dimensional vector that represents an observed frequency.

Let the total latent variables corresponding to $\boldsymbol{X}$ be similarly represented by $\boldsymbol{Z} = \{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_D\}$, where $\boldsymbol{Z}_d = \{\boldsymbol{z}_{d1}, \cdots, \boldsymbol{z}_{dN_d}\}$. $\boldsymbol{z}_{dn}$ is a KM-dimensional vector in which only one entry, z_{dnkm}, takes a value of 1 and the others take values of 0 when frequency $\boldsymbol{x}_{dn}$ is generated from partial m $(1 \leq m \leq M)$ of basis k $(1 \leq k \leq K)$.

3. COMPARISON WITH RELATED WORK

The properties of iLHA are intermediate between those of two successful approaches–statistical inference and NMF–which are discussed here for comparison and to clarify the positioning of our approach.

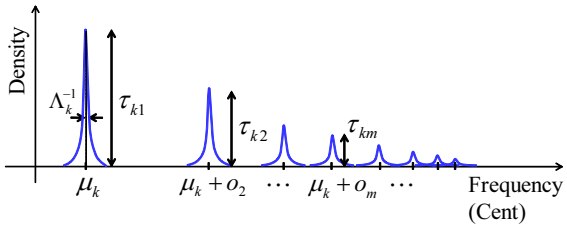

Figure 2. Probabilistic model of a single basis.

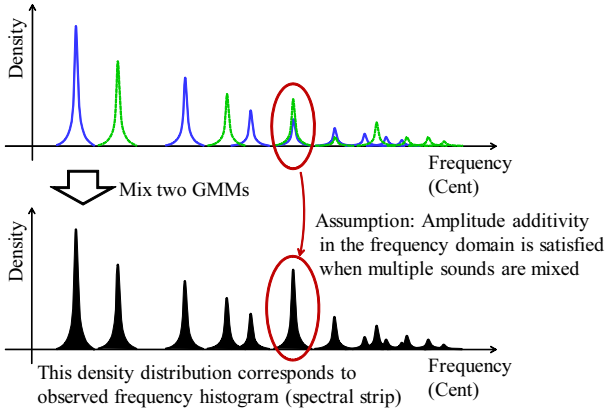

Figure 3. Probabilistic model of mixed multiple bases.

3.1 Statistical Inference

Statistical methods [3–6] assume probabilistic models using a limited number of parameters to represent the generative process of observed spectra (audio signals). F0 estimation directly corresponds to finding model parameters that provide the best explanations of the given data.

Goto [3] first proposed probabilistic models of harmonic sounds by regarding frequency spectra as probabilistic densities (histograms of observed frequencies).

As shown in Figure 2, the spectral distribution of basis k $(1 \leq k \leq K)$ is modeled by a harmonic GMM:

$$\mathcal{M}_k(\boldsymbol{x}) = \sum_{m=1}^{M} \tau_{km} \mathcal{N}\left(\boldsymbol{x} \middle| \boldsymbol{\mu}_k + \boldsymbol{o}_m, \boldsymbol{\Lambda}_k^{-1}\right), \qquad (1)$$

where $\boldsymbol{x}$ is a one-dimensional vector that indicates an observed frequency [cents]. [1] The Gaussian parameters, mean $\boldsymbol{\mu}_k$ and precision $\boldsymbol{\Lambda}_k$, indicate F0 [cents] of basis k and a degree of energy concentration around the F0 in the frequency domain. τ_{km} is a relative strength of the m-th harmonic partial $(1 \leq m \leq M)$ in basis k. We set $\boldsymbol{o}_m$ to $[1200 \log_2 m]$. This means M Gaussians are located to have harmonic relationships on the logarithmic frequency scale.

As shown in Figure 3, the spectral strip of frame d is modeled by mixing K harmonic GMMs as follows:

$$\mathcal{M}_d(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_{dk} \mathcal{M}_k(\boldsymbol{x}) \qquad (2)$$

where π_{dk} is a relative strength of basis k in frame d. Therefore, the polyphonic spectral strip is represented by nested finite Gaussian mixture models.

Several inference methods that have been proposed for parameter estimation are listed in Table 1. Goto [3] pro-

[1] Linear frequency f_h in hertz can be converted to logarithmic frequency f_c in cents as $f_c = 1200 \log_2(f_h/(440^{\frac{3}{12}-5}))$.

	#(bases)	#(partials)	Temporal modeling
PreFEst [3]	Fixed	Fixed	None
HC [4]	Inferred	Fixed	None
HTC [5]	Fixed	Fixed	Continuity treated
NMF [7]	Fixed	Not used	Exchangeable
iLHA	Infinite	Infinite	Exchangeable

Table 1. Comparison of multipitch analysis methods.

posed a method called PreFEst that estimates only relative strengths $\boldsymbol{\tau}$ and $\boldsymbol{\pi}$ while $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ are fixed by allocating many GMMs to cover the entire frequency range as F0 candidates. Kameoka *et al.* [4] then proposed harmonic clustering (HC), which estimates all the parameters and selects the optimal number of bases by using the Akaike information criterion (AIC). Although these methods yielded the promising results, they analyze the spectral strips of different frames independently. Thus, Kameoka *et al.* [5] proposed harmonic-temporal-structured clustering (HTC) that captures temporal continuity of spectral bases. Note that all these methods are based on maximum-likelihood and maximum-a-posteriori training of the parameters by introducing prior distributions of relative strengths $\boldsymbol{\tau}$, which have a strong impact on the accuracy of F0 estimation.

Our method called iLHA is based on hierarchical nonparametric Bayesian modeling that requires no prior tuning and avoids specifying K and M in advance. More specifically, the limit of the conventional nested finite GMMs is considered as K and M diverge to infinity.

3.2 Nonnegative Matrix Factorization

NMF-based methods [7–12] factorize observed frequency spectra into the product of spectral bases and time-varying envelopes under the nonnegativity constraint. K bases are estimated by sweeping all frames of the given spectra. Although several methods [10, 12] take temporal continuity into account, standard methods are based on temporal exchangeability. In other words, exchange of arbitrary frames does not affect the factorized results. Although such temporal modeling is not sufficient, it is known to work well in practice. Therefore, iLHA adopted the exchangeability.

4. LATENT HARMONIC ALLOCATION

This section explains LHA, the finite version of iLHA, as a preliminary step to deriving iLHA. We formulate the conventional nested *finite* GMMs in a Bayesian manner.

4.1 Model Formulation

Figure 4 illustrates a graphical representation of the LHA model. The full joint distribution is given by

$$\begin{aligned} &p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda}) \\ &= p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\boldsymbol{Z}|\boldsymbol{\pi},\boldsymbol{\tau})p(\boldsymbol{\pi})p(\boldsymbol{\tau})p(\boldsymbol{\mu},\boldsymbol{\Lambda}) \end{aligned} \tag{3}$$

where the first two terms on the right-hand side are likelihood functions and the other three terms are prior distributions. The likelihood functions are defined as

$$p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{dnkm} \mathcal{N}\left(\boldsymbol{x}_{dn}\middle|\boldsymbol{\mu}_k + \boldsymbol{o}_m, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{dnkm}} \tag{4}$$

$$p(\boldsymbol{Z}|\boldsymbol{\pi},\boldsymbol{\tau}) = \prod_{dnkm} (\pi_{dk}\tau_{km})^{z_{dnkm}} \tag{5}$$

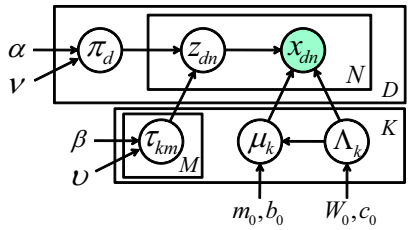

Figure 4. A graphical representation of LHA.

Then, we introduce conjugate priors as follows:

$$p(\boldsymbol{\pi}) = \prod_{d=1}^{D} \mathrm{Dir}(\boldsymbol{\pi}_d|\alpha\boldsymbol{\nu}) \propto \prod_{d=1}^{D}\prod_{k=1}^{K} \pi_{dk}^{\alpha\nu_k - 1} \tag{6}$$

$$p(\boldsymbol{\tau}) = \prod_{k=1}^{K} \mathrm{Dir}(\boldsymbol{\tau}_k|\beta\boldsymbol{\upsilon}) \propto \prod_{k=1}^{K}\prod_{m=1}^{M} \tau_{km}^{\beta\upsilon_m - 1} \tag{7}$$

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k\middle|\boldsymbol{m}_0,(b_0\boldsymbol{\Lambda}_k)^{-1}\right)\mathcal{W}\left(\boldsymbol{\Lambda}_k\middle|\boldsymbol{W}_0,c_0\right) \tag{8}$$

where $p(\boldsymbol{\pi})$ and $p(\boldsymbol{\tau})$ are products of Dirichlet distributions and $p(\boldsymbol{\mu},\boldsymbol{\Lambda})$ is a product of Gaussian-Wishart distributions. $\alpha\boldsymbol{\nu}$ and $\beta\boldsymbol{\upsilon}$ are hyperparameters and α and β are called concentration parameters when $\boldsymbol{\nu}$ and $\boldsymbol{\upsilon}$ sum to unity. $\boldsymbol{m}_0$, b_0, $\boldsymbol{W}_0$, and c_0 are also hyperparameters; $\boldsymbol{W}_0$ is a scale matrix and c_0 is a degree of freedom.

4.2 Variational Bayesian Inference

The objective of Bayesian inference is to compute a true posterior distribution of all variables: $p(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{X})$. Because analytical calculation of the posterior distribution is intractable, we instead approximate it by using iterative inference techniques such as variational Bayes (VB) and Markov chain Monte Carlo (MCMC). Although MCMC is considered to be more accurate in general, we use VB because it converges much faster.

In the VB framework, we introduce a variational posterior distribution $q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda})$ and make it close to the true posterior $p(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{X})$ iteratively. Here, we assume that the variational distribution can be factorized as

$$q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda}) = q(\boldsymbol{Z})q(\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda}) \tag{9}$$

To optimize $q(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda})$, we use a variational version of the Expectation-Maximization (EM) algorithm [13]. We iterate VB-E and VB-M steps until a variational lower bound of evidence $p(\boldsymbol{X})$ converges as follows:

$$q^*(\boldsymbol{Z}) \propto \exp\left(\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[\log p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda})\right]\right) \tag{10}$$

$$q^*(\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda}) \propto \exp\left(\mathbb{E}_{\boldsymbol{Z}}\left[\log p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda})\right]\right) \tag{11}$$

4.3 Updating Formula

We derive the formulas for updating variational posterior distributions according to Eqns. (10) and (11).

4.3.1 VB-E Step

An optimal variational posterior distribution of latent variables $\boldsymbol{Z}$ can be computed as follows:

$$\begin{aligned} &\log q^*(\boldsymbol{Z}) = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[\log p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda})\right] + \text{const.} \\ &= \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[\log p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})\right] + \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\tau}}\left[\log p(\boldsymbol{Z}|\boldsymbol{\pi},\boldsymbol{\tau})\right] + \text{const.} \\ &= \sum_{dnkm} z_{dnkm}\log\rho_{dnkm} + \text{const.} \end{aligned} \tag{12}$$

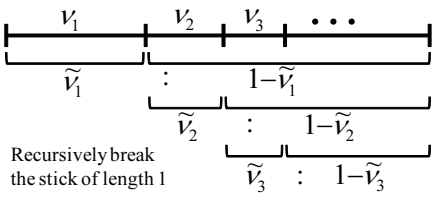

Figure 5. Stick-breaking construction of Dirichlet process.

where ρ_{dnkm} is defined as

$$\log \rho_{dnkm} = \mathbb{E}_{\boldsymbol{\pi}_d}[\log \pi_{dk}] + \mathbb{E}_{\boldsymbol{\tau}_k}[\log \tau_{km}] + \mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[\log \mathcal{N}\left(\boldsymbol{x}_{dn}\middle|\boldsymbol{\mu}_k + \boldsymbol{o}_m, \boldsymbol{\Lambda}_k^{-1}\right)\right] \quad (13)$$

$q^*(\boldsymbol{Z})$ is obtained as multinomial distributions given by

$$q^*(\boldsymbol{Z}) = \prod_{dnkm} \gamma_{dnkm}^{z_{dnkm}} \quad (14)$$

where γ_{dnkm} is given by $\gamma_{dnkm} = \frac{\rho_{dnkm}}{\sum_{km}\rho_{dnkm}}$ and is called a responsibility that indicates how likely it is that observed frequency $\boldsymbol{x}_{dn}$ is generated from harmonic partial m of basis k. Here, let n_{dkm} be an observation count that indicates how many frequencies were generated from harmonic partial m of basis k in frame d. n_{dkm} and its expected value can be calculated as follows:

$$n_{dkm} = \sum_n z_{dnkm} \quad \mathbb{E}[n_{dkm}] = \sum_n \gamma_{dnkm} \quad (15)$$

For convenience in executing the VB-M step, we compute several sufficient statistics as follows:

$$\mathbb{S}_k[1] \equiv \sum_{dnm} \gamma_{dnkm} \quad \mathbb{S}_k[\boldsymbol{x}] \equiv \sum_{dnm} \gamma_{dnkm}\boldsymbol{x}_{dnm} \quad (16)$$

$$\mathbb{S}_k[\boldsymbol{x}\boldsymbol{x}^T] \equiv \sum_{dnm} \gamma_{dnkm}\boldsymbol{x}_{dnm}\boldsymbol{x}_{dnm}^T \quad (17)$$

where $\boldsymbol{x}_{dnm}$ is defined as $\boldsymbol{x}_{dnm} = \boldsymbol{x}_{dn} - \boldsymbol{o}_m$.

4.3.2 VB-M Step

Consequently, an optimal variational posterior distribution of parameters $\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$ is shown to be given by

$$q^*(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{d=1}^{D} q^*(\boldsymbol{\pi}_d) \prod_{k=1}^{K} q^*(\boldsymbol{\tau}_k) \prod_{k=1}^{K} q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \quad (18)$$

Since we use conjugate priors, each posterior has the same form of the corresponding prior as follows:

$$q^*(\boldsymbol{\pi}_d) = \mathrm{Dir}(\boldsymbol{\pi}_d|\boldsymbol{\alpha}_d) \quad (19)$$

$$q^*(\boldsymbol{\tau}_k) = \mathrm{Dir}(\boldsymbol{\tau}_k|\boldsymbol{\beta}_k) \quad (20)$$

$$q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}\left(\boldsymbol{\mu}_k\middle|\boldsymbol{m}_k, (b_k\boldsymbol{\Lambda}_k)^{-1}\right)\mathcal{W}\left(\boldsymbol{\Lambda}_k\middle|\boldsymbol{W}_k, c_k\right) \quad (21)$$

where the variational parameters are given by

$$\alpha_{dk} = \alpha\nu_k + \mathbb{E}[n_{dk\cdot}] \quad \beta_{km} = \beta\upsilon_m + \mathbb{E}[n_{\cdot km}] \quad (22)$$

$$b_k = b_0 + \mathbb{S}_k[1] \quad c_k = c_0 + \mathbb{S}_k[1] \quad (23)$$

$$\boldsymbol{m}_k = \frac{b_0\boldsymbol{m}_0 + \mathbb{S}_k[\boldsymbol{x}]}{b_0 + \mathbb{S}_k[1]} = \frac{b_0\boldsymbol{m}_0 + \mathbb{S}_k[\boldsymbol{x}]}{b_k} \quad (24)$$

$$\boldsymbol{W}_k^{-1} = \boldsymbol{W}_0^{-1} + b_0\boldsymbol{m}_0\boldsymbol{m}_0^T + \mathbb{S}_k[\boldsymbol{x}\boldsymbol{x}^T] - b_k\boldsymbol{m}_k\boldsymbol{m}_k^T \quad (25)$$

Here dot $(\cdot)$ denotes the sum over that index.

5. INFINITE LATENT HARMONIC ALLOCATION

This section derives hierarchical nonparametric Bayesian models, i.e., nested *infinite* GMMs for polyphonic spectra.

5.1 Model Formulation

First we let K approach infinity, where the infinite number of harmonic GMMs is assumed to exist in the universe. More specifically, the dimensionality of the Dirichlet distributions in Eqn. (6) is considered to be infinite. At each frame d, $\boldsymbol{\pi}_d$ is an infinite vector of normalized probabilities (mixing weights) drawn from the infinite-dimensional Dirichlet prior. Such stochastic process is called a Dirichlet process (DP). Every time frequency $\boldsymbol{x}_{dn}$ is generated, one of the infinite number of harmonic GMMs is drawn according to $\boldsymbol{\pi}_d$. Note that most entries of $\boldsymbol{\pi}_d$ take extremely tiny values because all entries sum to unity. If we can observe the infinite number of frequencies ($N_d \to \infty$), the infinite number of harmonic GMMs can be drawn. However, N_d is finite in practice. Therefore, only the finite number of harmonic GMMs, $K_+ \ll \infty$, is drawn at frame d. Here, a problem is that harmonic GMMs that are actually drawn at frame d are completely disjointed from those drawn at another frame d'. This is not a reasonable situation.

To solve this problem, we use the hierarchical Dirichlet Process (HDP) [14]. More specifically, we assume that infinite-dimensional hyperparameter $\boldsymbol{\nu}$ in Eqn. (6), which is shared among all D frames, is a draw from a top-level DP. A generative interpretation is that after an unbounded number of harmonic GMMs is initially drawn from the top-level DP, an unbounded subset is further drawn according to the local DP at each frame. This effectively ties frame d to another frame d'. As shown in Figure 5, $\boldsymbol{\nu}$ is known to follow the stick-breaking construction [14] as follows:

$$\nu_k = \tilde{\nu}_k \prod_{k'=1}^{k-1}(1 - \tilde{\nu}_{k'}) \quad \tilde{\nu}_k \sim \mathrm{Beta}(1, \gamma) \quad (26)$$

where γ is a concentration parameter of the top-level DP. Therefore $\boldsymbol{\nu}$ can be converted into $\tilde{\boldsymbol{\nu}}$.

Now we let M approach infinity, where each harmonic GMM consists of the infinite number of harmonic partials. To put effective priors on $\boldsymbol{\tau}$, we use generalized DPs called Beta two-parameter processes as follows:

$$\tau_{km} = \tilde{\tau}_{km} \prod_{m'=1}^{m-1}(1 - \tilde{\tau}_{km'}) \quad \tilde{\tau}_{km} \sim \mathrm{Beta}(\beta\lambda_1, \beta\lambda_2) \quad (27)$$

where β is a positive scalar and $\lambda_1 + \lambda_2 = 1$.

Because α, β, γ and $\boldsymbol{\lambda}$ are influential hyperparameters, we put Gamma and Beta hyperpriors on them as follows:

$$p(\alpha) = \mathrm{Gam}(\alpha|a_\alpha, b_\alpha) \quad p(\gamma) = \mathrm{Gam}(\gamma|a_\gamma, b_\gamma) \quad (28)$$

$$p(\beta) = \mathrm{Gam}(\beta|a_\beta, b_\beta) \quad p(\boldsymbol{\lambda}) = \mathrm{Beta}(\boldsymbol{\lambda}|u_1, u_2) \quad (29)$$

where $a_{\{\alpha,\beta,\gamma\}}$ and $b_{\{\alpha,\beta,\gamma\}}$ are shape and rate parameters.

Figure 6 shows a graphical representation of the iLHA model. The full joint distribution is given by

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \tilde{\boldsymbol{\tau}}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha, \beta, \gamma, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}) = p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ p(\boldsymbol{Z}|\boldsymbol{\pi}, \tilde{\boldsymbol{\tau}})p(\boldsymbol{\pi}|\alpha, \tilde{\boldsymbol{\nu}})p(\tilde{\boldsymbol{\tau}}|\beta, \boldsymbol{\lambda})p(\alpha)p(\beta)p(\gamma)p(\boldsymbol{\lambda})p(\tilde{\boldsymbol{\nu}}|\gamma) \quad (30)$$

where $p(\boldsymbol{Z}|\boldsymbol{\pi}, \tilde{\boldsymbol{\tau}})$ is obtained by plugging Eqn. (27) into Eqn. (5) and $p(\boldsymbol{\pi}|\alpha, \boldsymbol{\nu})$ is the same as Eqn. (6). $p(\tilde{\boldsymbol{\nu}}|\gamma)$ and $p(\tilde{\boldsymbol{\tau}}|\beta, \boldsymbol{\lambda})$ are defined according to Eqns. (26) and (27) as

$$p(\tilde{\boldsymbol{\nu}}|\gamma) = \prod_k \mathrm{Beta}(\tilde{\tau}_k|1, \gamma) \; p(\tilde{\boldsymbol{\tau}}|\beta, \boldsymbol{\lambda}) = \prod_{km} \mathrm{Beta}(\tilde{\tau}_{km}|\beta\boldsymbol{\lambda}) \quad (31)$$

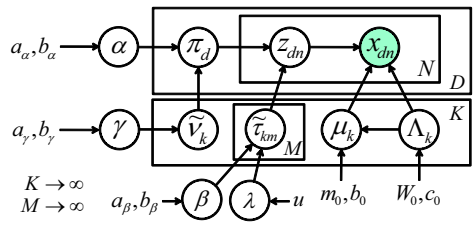

Figure 6. A graphical representation of iLHA.

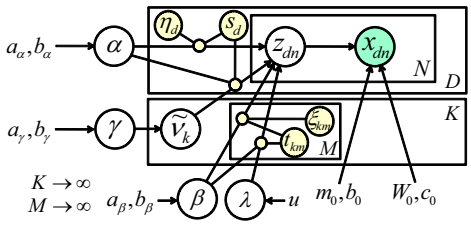

Figure 7. A collapsed model with auxiliary variables.

5.2 Collapsed Variational Bayesian Inference

To train the HDP model we use a sophisticated version of VB called collapsed variational Bayes (CVB) [15]. CVB enables more accurate posterior approximation in the space of latent variables where parameters are integrated out.

Figure 7 shows a collapsed iLHA model. By integrating out $\boldsymbol{\pi}, \tilde{\boldsymbol{\tau}}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$, we obtain the marginal distribution given by

$$p(\boldsymbol{X}, \boldsymbol{Z}, \alpha, \beta, \gamma, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}) = p(\boldsymbol{X}|\boldsymbol{Z})p(\boldsymbol{Z}|\alpha, \beta, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}})p(\alpha)p(\beta)p(\gamma)p(\boldsymbol{\lambda})p(\tilde{\boldsymbol{\nu}}|\gamma) \quad (32)$$

where the first two terms are calculated as follows:

$$p(\boldsymbol{X}|\boldsymbol{Z}) = (2\pi)^{-\frac{n_{\cdots}}{2}} \prod_k \left(\frac{b_0}{b_{zk}}\right)^{\frac{1}{2}} \frac{B(\boldsymbol{W}_0, c_0)}{B(\boldsymbol{W}_{zk}, c_{zk})} \quad (33)$$

$$p(\boldsymbol{Z}|\alpha, \beta, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}) = \prod_d \frac{\Gamma(\alpha)}{\Gamma(\alpha + n_{d\cdot\cdot})} \prod_k \frac{\Gamma(\alpha\nu_k + n_{dk\cdot})}{\Gamma(\alpha\nu_k)} \prod_{km} \frac{\Gamma(\beta)\Gamma(\beta\lambda_1 + n_{\cdot km})\Gamma(\beta\lambda_2 + n_{\cdot k>m})}{\Gamma(\beta\lambda_1)\Gamma(\beta\lambda_2)\Gamma(\beta + n_{\cdot k\geq m})} \quad (34)$$

where $b_{zk}, \boldsymbol{W}_{zk}, c_{zk}$ are obtained by substituting z_{dnkm} for γ_{dnkm} in calculating Eqns. (23) and (25).

Because CVB cannot be applied directly to Eqn. (32), we introduce auxiliary variables by using a technique called data augmentation [15]. Let η_d and ξ_{km} be Beta-distributed variables and s_{dk} and $\boldsymbol{t}_{km}$ be positive integers that satisfy $1 \leq s_{dk} \leq n_{dk\cdot}$, $1 \leq t_{km1} \leq n_{\cdot km}$, and $1 \leq t_{km2} \leq n_{\cdot k>m}$. Eqn. (34) can be augmented as

$$p(\boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{s}, \boldsymbol{t}|\alpha, \beta, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}) = \prod_d \frac{\eta_d^{\alpha-1}(1-\eta_d)^{n_{d\cdot\cdot}-1}}{\Gamma(n_{d\cdot\cdot})} \prod_k \begin{bmatrix} n_{dk\cdot} \\ s_{dk} \end{bmatrix} (\alpha\nu_k)^{s_{dk}} \prod_{km} \frac{\xi_{km}^{\beta-1}(1-\xi_{km})^{n_{\cdot k\geq m}-1}}{\Gamma(n_{\cdot k\geq m})} \begin{bmatrix} n_{\cdot km} \\ t_{km1} \end{bmatrix} (\beta\lambda_1)^{t_{km1}} \begin{bmatrix} n_{\cdot k>m} \\ t_{km2} \end{bmatrix} (\beta\lambda_2)^{t_{km2}}$$

where $[\ ]$ denotes a Stirling number of the first kind. The augmented marginal distribution is given by

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{s}, \boldsymbol{t}, \alpha, \beta, \gamma, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}) = p(\boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{s}, \boldsymbol{t}|\alpha, \beta, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}})p(\alpha)p(\beta)p(\gamma)p(\boldsymbol{\lambda})p(\tilde{\boldsymbol{\nu}}|\gamma) \quad (35)$$

In the CVB framework, we assume that the variational posterior distribution can be factorized as follows:

$$q(\boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{s}, \boldsymbol{t}, \alpha, \beta, \gamma, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}) = q(\alpha, \beta, \gamma, \boldsymbol{\lambda})q(\tilde{\boldsymbol{\nu}})q(\boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{s}, \boldsymbol{t}|\boldsymbol{Z}) \prod_{dn} q(\boldsymbol{z}_{dn}) \quad (36)$$

We also use an approximation technique called variational posterior truncation. More specifically, we assume $q(z_{dnkm}) = 0$ when $k > K$ and $m > M$. In practice, it is enough that K and M are set to sufficiently large integers.

5.3 Updating Formula

We descrive the formulas for updating variational posterior distributions.

5.3.1 CVB-E Step

A variational probability of $z_{dnkm} = 1$ is given by

$$\log q^*(z_{dnkm} = 1) = \mathbb{E}_{\boldsymbol{z}^{\neg dn}}\left[\log\left(\mathbb{G}[\alpha\nu_k] + n_{dk\cdot}^{\neg dn}\right)\right] + \mathbb{E}_{\boldsymbol{z}^{\neg dn}}\left[\log\left(\frac{\mathbb{G}[\beta\lambda_1] + n_{\cdot km}^{\neg dn}}{\mathbb{E}[\beta] + n_{\cdot k\geq m}^{\neg dn}} \prod_{m'=1}^{m-1} \frac{\mathbb{G}[\beta\lambda_2] + n_{\cdot k>m'}^{\neg dn}}{\mathbb{E}[\beta] + n_{\cdot k\geq m'}^{\neg dn}}\right)\right] + \mathbb{E}_{\boldsymbol{z}^{\neg dn}}\left[\log \mathcal{S}(\boldsymbol{x}_{dnm}|\boldsymbol{m}_{zk}^{\neg dn}, \boldsymbol{L}_{zk}^{\neg dn}, c_{zk}^{\neg dn})\right] + \text{const.} \quad (37)$$

where subscript $\neg dn$ denotes a set of indices without d and n, $\mathbb{G}[x]$ denotes the geometric average $\exp(\mathbb{E}[\log x])$, and $\mathcal{S}$ is the Student-t distribution. $\boldsymbol{L}_{zk}^{\neg dn}$ is given by $\boldsymbol{L}_{zk}^{\neg dn} = \frac{b_{zk}^{\neg dn}}{1+b_{zk}^{\neg dn}} c_{zk}^{\neg dn} \boldsymbol{W}_{zk}^{\neg dn}$, where $\boldsymbol{m}_{zk}^{\neg dn}, b_{zk}^{\neg dn}, \boldsymbol{W}_{zk}^{\neg dn}, c_{zk}^{\neg dn}$ are obtained by substituting z_{dnkm} for γ_{dnkm} required by Eqns. (23), (24), and (25) and calculating sum without $\boldsymbol{z}_{dn}$. Each term of Eqn. (37) can be calculated efficiently [15, 16].

5.3.2 CVB-M Step

First, α, β and γ are Gamma distributed as follows:

$$q(\alpha) \propto \alpha^{a_\alpha + \mathbb{E}[s_{\cdot\cdot}]-1} e^{-\alpha\left(b_\alpha - \sum_d \mathbb{E}[\log \eta_d]\right)} \quad (38)$$

$$q(\beta) \propto \beta^{a_\beta + \mathbb{E}[t_{\cdots}]-1} e^{-\beta\left(b_\beta - \sum_{km} \mathbb{E}[\log \xi_{km}]\right)} \quad (39)$$

$$q(\gamma) \propto \gamma^{a_\gamma + K - 1} e^{-\gamma\left(b_\gamma - \sum_k \mathbb{E}[\log(1-\tilde{\nu}_k)]\right)} \quad (40)$$

Then, $\boldsymbol{\lambda}$ and $\tilde{\boldsymbol{\tau}}$ are Beta distributed as follows:

$$q^*(\boldsymbol{\lambda}) \propto \lambda_1^{u_1 + \mathbb{E}[t_{\cdot\cdot 1}]-1} \lambda_2^{u_2 + \mathbb{E}[t_{\cdot\cdot 2}]-1} \quad (41)$$

$$q^*(\tilde{\nu}_k) \propto \tilde{\nu}_k^{1 + \mathbb{E}[s_{\cdot k}]-1} (1-\tilde{\nu}_k)^{\mathbb{E}[\gamma] + \mathbb{E}[s_{\cdot >k}]-1} \quad (42)$$

Finally, the variational posteriors of $\boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{s}, \boldsymbol{t}$ are given by

$$q^*(\eta_d) \propto \eta_d^{\mathbb{E}[\alpha]-1}(1-\eta_d)^{n_{d\cdot\cdot}-1} \quad (43)$$

$$q^*(\xi_{km}|\boldsymbol{Z}) \propto \xi_{km}^{\mathbb{E}[\beta]-1}(1-\xi_{km})^{n_{\cdot k\geq m}-1} \quad (44)$$

$$q^*(s_{dk} = s|\boldsymbol{Z}) \propto \begin{bmatrix} n_{dk\cdot} \\ s \end{bmatrix} \mathbb{G}[\alpha\nu_k]^s \quad (45)$$

$$q^*(t_{km1} = t|\boldsymbol{Z}) \propto \begin{bmatrix} n_{\cdot km} \\ t \end{bmatrix} \mathbb{G}[\beta\lambda_1]^t \quad (46)$$

$$q^*(t_{km2} = t|\boldsymbol{Z}) \propto \begin{bmatrix} n_{\cdot k>m} \\ t \end{bmatrix} \mathbb{G}[\beta\lambda_2]^t \quad (47)$$

To calculate $\mathbb{E}[s_{dk}]$ (average s_{dk} over $\boldsymbol{Z}$), we exactly treat the case $n_{dk\cdot} = 0$ and apply second-order approximation when $n_{dk\cdot} > 0$ (see details in [15]). $\mathbb{E}[\log \xi_{km}]$, $\mathbb{E}[t_{km1}]$, and $\mathbb{E}[t_{km2}]$ can be calculated in the same way.

To estimate F0s, we need explicitly compute the variational posteriors of the integrated-out parameters $\boldsymbol{\mu}, \boldsymbol{\Lambda}$. To do this, we execute the standard VB-M step once by using the responsibilities $q(\boldsymbol{Z})$ obtained in the CVB-E step.

6. EVALUATION

This section reports our comparative experiments evaluating the performance of iLHA.

Piece number	Optimally tuned		**Fully automated**	
RWC-MDB-	PreFEst [3]	HTC [5]	LHA	iLHA
J-2001 No.1	75.8	79.0	70.7	**82.2**
J-2001 No.2	**78.5**	78.0	69.1	77.9
J-2001 No.6	70.4	**78.3**	49.8	71.2
J-2001 No.7	83.0	**86.0**	70.2	85.5
J-2001 No.8	**85.7**	84.4	55.9	84.6
J-2001 No.9	85.9	**89.5**	68.9	84.7
C-2001 No.30	76.0	**83.6**	81.4	81.6
C-2001 No.35	72.8	76.0	58.9	**79.6**
Total	79.4	**82.0**	65.8	81.7

Table 2. Frame-level F-measures of F0 detection.

6.1 Experimental Conditions

We evaluated LHA and iLHA on the same test set used in [5], which consisted of nine pieces of piano and guitar solo performances excerpted from the RWC music database [17]. The first 23 [s] of each piece were used for evaluation. The spectral analysis was conducted by the wavelet transform using Gabor wavelets with a time resolution of 16 [ms]. The values and temporal positions of actual F0s were prepared by hand as ground truth. We evaluated performance in terms of frame-level F-measures. The priors and hyperpriors of LHA and iLHA were set to noninformative uniform distributions. K and M were set to sufficiently large numbers, 60 and 15. iLHA is not sensitive to these values. No other tuning was required. To output F0s at each frame, we extracted bases whose expected weights $\boldsymbol{\pi}$ were over a threshold, which was optimized as in [5].

For comparison, we referred to the experimental results of PreFEst and HTC reported in [5]. Although the ground-truth data was slightly different from ours, it would be sufficient for roughly evaluating performance comparatively. The number of bases, priors, and weighting factors were carefully tuned by using the ground-truth data to optimize the results. Although this is not realistic, the *upper bounds* of potential performance were investigated in [5].

6.2 Experimental Results

The results listed in Table 2 show that the performance of iLHA approached and sometimes surpassed that of HTC. This is consistent with the empirical findings of many studies on Bayesian nonparametrics that nonparametric models were competitive against optimally-tuned parametric models. HTC outperformed PreFEst because HTC can appropriately deal with temporal continuity of spectral bases. This implies that incorporating temporal modeling would improve the performance of iLHA.

The results of LHA were worse than those of iLHA because LHA is not based on hierarchical Bayesian modeling and requires precise priors. In fact, we confirmed that the results of PreFEst and HTC based on MAP estimation were drastically degraded when we used noninformative priors. In contrast, iLHA stably showed the good performance.

7. CONCLUSION

This paper presented a novel statistical method for detecting multiple F0s in polyphonic audio signals. The method allows polyphonic spectra to contain an unbounded number of spectral bases, each of which can consist of an unbounded number of harmonic partials. These numbers can be statistically inferred at the same time that F0s are estimated. Even in experimental evaluation using noninformative priors, our automated method performed well or better than conventional methods manually optimized by trial and error. To our knowledge, this is the first attempt to apply Bayesian nonparametrics to multipitch analysis.

Bayesian nonparametrics is an ultimate methodological framework avoiding the model selection problem faced in various areas of MIR. For example, how many sections should one use for structuring a musical piece? How many groups should one use for clustering listeners according to their tastes or musical pieces according to their contents? We are freed from these problems by assuming that in theory there is an infinite number of classes behind observed data. Unnecessary classes are automatically removed from consideration through statistical inference. We plan to use this powerful framework in a wide range of applications.

Acknowledgement: This study was partially supported by JST CREST and JSPS KAKENHI 20800084.

8. REFERENCES

[1] A. Klapuri. Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Trans. on ASLP*, Vol. 16, No. 2, pp. 255–266, 2008.

[2] M. Marolt. A connectionist approach to transcription of polyphonic piano music. *IEEE Trans. on Multimedia*, Vol. 6, No. 3, pp. 439–449, 2004.

[3] M. Goto. A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, Vol. 43, No. 4, pp. 311–329, 2004.

[4] H. Kameoka *et al*. Separation of harmonic structures based on tied Gaussian mixture model and information criterion for concurrent sounds. *ICASSP*, Vol. 4, pp. 297–300, 2004.

[5] H. Kameoka *et al*. A multipitch analyzer based on harmonic temporal structured clustering. *IEEE Trans. on ASLP*, Vol. 15, No. 3, pp. 982–994, 2007.

[6] A. Cemgil *et al*. A generative model for music transcription. *IEEE Trans. on ASLP*, Vol. 14, No. 2, pp. 679–694, 2006.

[7] P. Smaragdis and J. Brown. Nonnegative matrix factorization for polyphonic music transcription. *WASPAA*, 2003.

[8] A. Cont. Realtime multiple pitch observation using sparse non-negative constraints. *ISMIR*, pp. 206–211, 2006.

[9] E. Vincent *et al*. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Trans. on ASLP*, Vol. 18, No. 3, pp. 528–537, 2010.

[10] N. Bertin *et al*. Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Trans. on ASLP*, Vol. 18, No. 3, pp. 538–549, 2010.

[11] P. Peeling *et al*. Generative spectrogram factorization models for polyphonic piano transcription. *IEEE Trans. on ASLP*, Vol. 18, No. 3, pp. 519–527, 2010.

[12] T. Virtanen *et al*. Bayesian extensions to nonnegative matrix factorisation for audio signal modelling. *ICASSP*, 2008.

[13] H. Attias. A variational Bayesian framework for graphical models. *NIPS*, pp. 209–215, 2000.

[14] Y. W. Teh *et al*. Hierarchical Dirichlet processes. *J. of Am. Stat. Assoc.*, Vol. 101, No. 476, pp. 1566–1581, 2006.

[15] Y. W. Teh *et al*. Collapsed variational inference for HDP. *NIPS*, Vol. 20, 2008.

[16] J. Sung *et al*. Latent-space variational Bayes. *IEEE Trans. on PAMI*, Vol. 30, No. 12, pp. 2236–2242, 2008.

[17] M. Goto *et al*. RWC music database: Popular, classical, and jazz music database. *ISMIR*, pp. 287–288, 2002.