AUDIO STREAM SEGREGATION OF MULTI-PITCH MUSIC SIGNAL BASED ON TIME-SPACE CLUSTERING USING GAUSSIAN KERNEL 2-DIMENSIONAL MODEL

Hirokazu Kameoka, Takuya Nishimoto and Shigeki Sagayama

Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{kameoka,nishi,sagayama}@hil.t.u-tokyo.ac.jp

ABSTRACT

This paper describes a novel approach for audio stream segregation of multi-pitch music signal. We propose parameter-constrained time-frequency spectrum model expressing both harmonic spectral structure and temporal curve of power envelope with Gaussian kernels. MAP estimation of the model parameters using EM algorithm provides fundamental frequency, onset and offset time, spectral envelope and power envelope of every underlying audio stream. Our proposed method showed high accuracy in pitch name estimation task of several pieces of real music performance data.

1. INTRODUCTION

Multi-pitch audio signal analysis has been one of the important subjects in speech processing and music processing areas. Automatic music transcription and signal-to-MIDI conversion technique have been expected to be useful for music information retrieval purpose, which is one of the most attractive issues in the recent music processing area. These transformation works as a data compression allowing fast and flexible query search in the large existing music content database. Sound source separation has been, as always, a big concern in many research areas, e.g., robust speech recognition, audio coding and others.

Contrary to its high demands, however, the standard level of the numerous conventional methods has been far from a practical step. Yet the recent novel ideas, e.g., filter-bank approach[1], Kalman filtering based approach[2], signal-level and spectrum-level model approximation approach [3] [4], brought remarkable progress. While multi-pitch analysis is basically an ill-posed problem of finding likely solutions in time-frequency space, these methods made the problem solvable by dealing with each dimension separately: first extract accurate frequency-dimension information (e.g., pitch or pitch likelihood) from each short time segment and then give overall solution in long time interval by combining every local information together.

Apart from the common attempts based on dynamic integration of all subsequent local information in frequency-dimension, that often blinds us to the global perspective of time-frequency structure, our proposed method tries to find optimal solutions directly throughout the time-frequency (2-dimensional) plane.

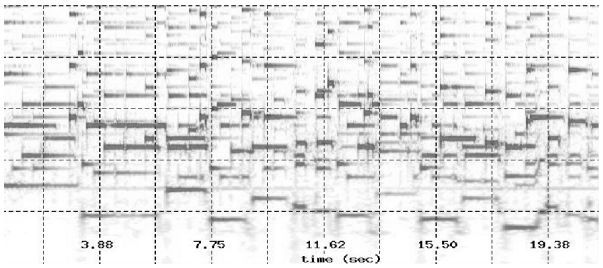


Fig. 1. Spectrogram of music signal ranging from T_0 to T_1 in time direction and from Ω_0 to Ω_1 in frequency direction

2. GAUSSIAN KERNEL AUDIO STREAM MODEL

2.1. Problem Formulation

As shown in Fig.1, multi-pitch music signal is a complex mixture of multiple audio streams each of which is associated with an musical note event. Segregating the mixture distribution into each spectrum is hardly a straightforward problem mainly because of spectral overlapping caused by spectral widening phenomena in short-time analysis. We adopt fuzzy-clustering-based approach for decomposing time-frequency plane into multiple striped-territories each of which occupies prospective spectral components of a single particular audio stream, under consideration that spectral distribution is a sort of histogram of "micro-energy" patterns.

Provided that each cluster assumes a geometric distribution $p(x,t|\mathbf{\Theta}_k)$ modeling power spectrum of a single audio stream (see Fig.2) determined by parameter $\mathbf{\Theta}_k$ ($\mathbf{\Theta} = \{\mathbf{\Theta}_k | k = 1, \cdots, K\}$), a particular form of the objective function for this clustering is given as

$$\sum_{k=1}^{K} \int_{T_0}^{T_1} \int_{\Omega_0}^{\Omega_1} \Big(p(k|x,t,\mathbf{\Theta}) f(x,t) \Big) \times D(x,t|\mathbf{\Theta}_k) dx dt \quad (1)$$

where x, t and $f(x,t)$ are log-frequency, time(frame) and spectral density of wavelet transform spectrum, T_0, T_1 and Ω_0, Ω_1 are the lower and higher bounds of time(frame) and log-frequency ranges, k and K are the index and the total

number of clusters, respectively. $p(k|x,t,\mathbf{\Theta})$ is a membership probability of the kth cluster at the coordinates (x,t), depending on every model parameter $\mathbf{\Theta}$, so that $p(k|x,t,\mathbf{\Theta})$ $f(x,t)$ means the spectral density of segregated audio stream. $D(x,t|\mathbf{\Theta}_k)$ is a measure function that suggests how dominant the kth model is at the coordinates (x,t). To put it more plainly, when the integral of the model density function has to be always equal to that of the given spectrum, i.e., a situation that $p(x,t|\mathbf{\Theta}_k)$ must satisfies

$$\sum_{k=1}^{K}\int_{T_0}^{T_1}\int_{\Omega_0}^{\Omega_1} p(x,t|\mathbf{\Theta}_k)dxdt = \int_{T_0}^{T_1}\int_{\Omega_0}^{\Omega_1} f(x,t)dxdt = F,$$

$(p(k|x,t,\mathbf{\Theta})f(x,t))D(x,t|\mathbf{\Theta}_k)$ takes greater value if the two distributions, $p(x,t|\mathbf{\Theta}_k)$ and $p(k|x,t,\mathbf{\Theta})f(x,t)$, get closer to each other. Hence approximating the given overall spectral distribution by the mixture of multiple audio stream models leads to maximizing Eq. (1). This equation is substantially same as Q function in EM algorithm particularly when $D(x,t|\mathbf{\Theta}_k) = \log p(x,t|\mathbf{\Theta}_k)$. The optimally-determined membership probability $p(k|x,t,\widehat{\mathbf{\Theta}})$ works as a spectral filter that only passes the target audio stream k.

In modeling the audio stream model $p(x,t|\mathbf{\Theta}_k)$, we should focus on 2 significant factors: harmonicity, and continuity of power envelope curve. In the following, we propose geometric model reflecting aforementioned 2 factors and discuss how the model parameters are estimated.

2.2. Model Description

Roughly assuming that pitch trajectory of a single audio stream is [1]parallel to the time axis (does not depend on t), a cutting plane of the kth audio stream model (Fig.2) at particular time t appears as Fig.3, which is a harmonic structure model $h_k(x)$ of fundamental log-frequency μ_k, weighted with power envelope curve function $g_k(t)$ (Fig.4). Hence the kth audio stream model $p(x,t|\mathbf{\Theta}_k)$ is simply expressed as a multiplication of the 2 functions and power w_k

$$p(x,t|\mathbf{\Theta}_k) = w_k h_k(t) g_k(x) \qquad (2)$$

$$\text{where}: \sum_{k=1}^{K} w_k = F, \quad \int_{\Omega_0}^{\Omega_1} h_k(x)dx = \int_{T_0}^{T_1} g_k(t)dt = 1.$$

Assuming ideal harmonicity, nth partial log-frequency is always located $\log n$ away from the fundamental log-frequency, so that, given the fundamenal log-frequency estimate μ_k, nth partial log-frequency estimate is μ_k+$\log n$. Now if each frequency component distribution can be approximated by a Guassian, a single harmonic structure can be modeled with a weighted sum of Gaussian kernels described as

$$h_k(x) = \sum_{n=1}^{N} \frac{r_n^k}{\sqrt{2\pi{\sigma_k}^2}} \exp\left[-\frac{\{x-(\mu_k+\log n)\}^2}{2{\sigma_k}^2}\right]. \qquad (3)$$

[1]This assumption was only made for simplifying the problem and does not limit the potential of our method. We shall leave further discussion of modeling pitch trajectory curve to our future work.

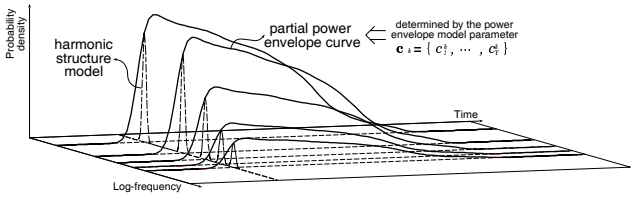


Fig. 2. Parametric audio stream model

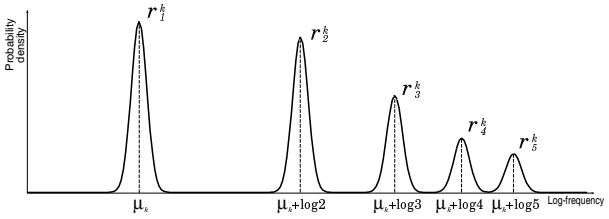


Fig. 3. Gaussian kernel harmonic structure model

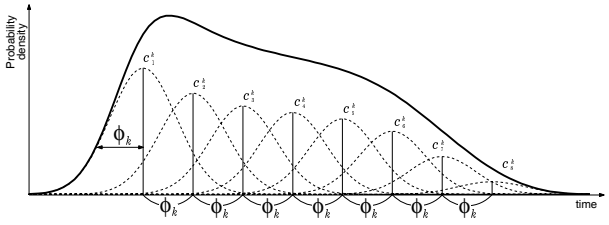


Fig. 4. Gaussian kernel power envelope model

where each weight parameter $r_n^k(\sum_{n=1}^{N} r_n^k = 1)$ is exactly related to the spectral components.

Since power envelopes may generally vary, depending on instruments or musical expressions, the curve function model $g_k(t)$ should be flexible enough to adapt well to every possible envelope. We formulate this model with Gaussian kernels each of which is weighted with $c_y^k(\sum_{y=0}^{Y-1} c_y^k = 1)$, that directly determines the shape of power envelope, where y is the index of the Gaussian. A specific feature of this model: the standard deviation of each Gaussian and the interval of adjacent Gaussians are expressed with a same variable ϕ_k, makes $g_k(t)$ a linear elastic function allowing various time lengths of audio streams. $g_k(t)$ is given as

$$g_k(t) = \sum_{y=0}^{Y-1} \frac{c_y^k}{\sqrt{2\pi{\phi_k}^2}} \exp\left[-\frac{\{t-(o_k+y\phi_k)\}^2}{2{\phi_k}^2}\right] \qquad (4)$$

where Y is the number of the Gaussian kernels and o_k is the center of the forefront Gaussian.

The whole parameters are listed in table2.2 together with the corresponding physical quantities or events.

3. ESTIMATING OPTIMAL PARAMETERS

3.1. *A Priori* Distribution

If we have a prior knowledge or an expectation of how spectral and power envelopes would shape like, *a priori* distribution assumption for r_n^k and c_y^k prevents the model from excessive deviation from the 'expected' envelopes (see Fig.5

Table 1. List of the free parameters of kth audio stream model

parameters	model	close physical corresopondence
μ_k	mean of the forefront Gaussian kernel in the harmonic structure model	fundamental log-frequency
$\mu_k + \log n$	mean of the nth Gaussian kernel in the harmonic structure model	nth partial log-frequency
w_k	weight	relative dominance of the kth audio stream
r_n^k	weights of Gaussian kernels in the harmonic structure model	spectral envelope
c_y^k	weights of Gaussian kernels in the power envelope model	temporal curve of power envelope
o_k	mean of the forefront Gaussian kernel in the power envelope model	onset time of the kth audio stream
σ_k	standard deviation of Gaussian kernel in the harmonic structure model	width of the frequency component
ϕ_k	interval & standard deviation of Gaussian kernels in the power envelope model	temoral length of the kth audio stream

for example). Here we apply the *a priori* distribution, proposed by Goto[4], to $\{r_n^k\}_{n=1}^N$ and $\{c_y^k\}_{y=0}^{Y-1}$ given by the exponential of nagative Kullback-Leibler distance between r_n^k, c_y^k and the 'expected' values $\bar{r}_n$, $\bar{c}_y$

$$p(\boldsymbol{r}_k) = \frac{1}{\beta(d_r)} \exp\Big(- d_r \sum_{n=1}^{N} \bar{r}_n \log \frac{\bar{r}_n}{r_n^k} \Big), \qquad (5)$$

$$p(\boldsymbol{c}_k) = \frac{1}{\beta(d_c)} \exp\Big(- d_c \sum_{y=0}^{Y-1} \bar{c}_y \log \frac{\bar{c}_y}{c_y^k} \Big). \qquad (6)$$

$$\sum_{n=1}^{N} r_n^k = \sum_{n=1}^{N} \bar{r}_n = 1, \quad \sum_{y=0}^{Y-1} c_y^k = \sum_{y=0}^{Y-1} \bar{c}_y = 1$$

where d_r and d_c are the influences of the *a priori* distributions, $\beta(d_r)$ and $\beta(d_c)$ are normalization factors. The advantage of using this particular form is in a considerable simplification of calculating Lagrange multipliers in MAP estimation without affecting its substance. Instead of this distribution, dirichlet distribution is also applicable.

3.2. MAP Estimation Using EM Algorithm

The conditional optimization problem of the time-frequency clustering using abovementioned paramater-constrained audio stream model has the same form of MAP(Maximum *A Posteriori*) estimation using EM algorithm. Since the objective function in Eq. (1) corresponds to the auxiliary function related to Q function, this can be rewritten as

$$\begin{aligned} R(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}) = \sum_{k=1}^{K} \Bigg\{ \sum_{n=1}^{N} \sum_{y=0}^{Y-1} \Bigg(\int_{T_1}^{T_2} \int_{\Omega_1}^{\Omega_2} p(k, n, y | x, t, \boldsymbol{\Theta}) \\ \times f(x,t) \log p(x, t | \widehat{\boldsymbol{\Theta}}_k) dx dt \Bigg) + \log p(\widehat{\boldsymbol{r}}_k) \\ + \log p(\widehat{\boldsymbol{c}}_k) - \lambda_r^{(k)} \Bigg(\sum_{n=1}^{N} \widehat{r}_n^k - 1 \Bigg) \\ - \lambda_c^{(k)} \Bigg(\sum_{y=0}^{Y-1} \widehat{c}_y^k - 1 \Bigg) \Bigg\} - \lambda_w \Bigg(\sum_{k=1}^{K} \widehat{w}_k - 1 \Bigg) \qquad (7) \end{aligned}$$

where $\lambda_r^{(k)}$, $\lambda_c^{(k)}$ and λ_w are Lagrange multipliers. Note that in this case, $f(x,t)$ must be a normalized density function and w_k must satisfy $\sum_{k=1}^{K} w_k = 1$.

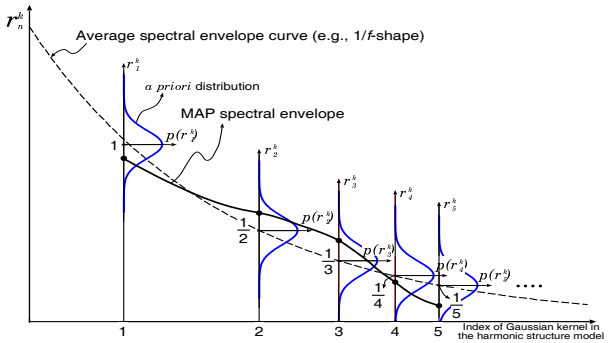


Fig. 5. *A priori* distribution of the weight parameter r_n^k

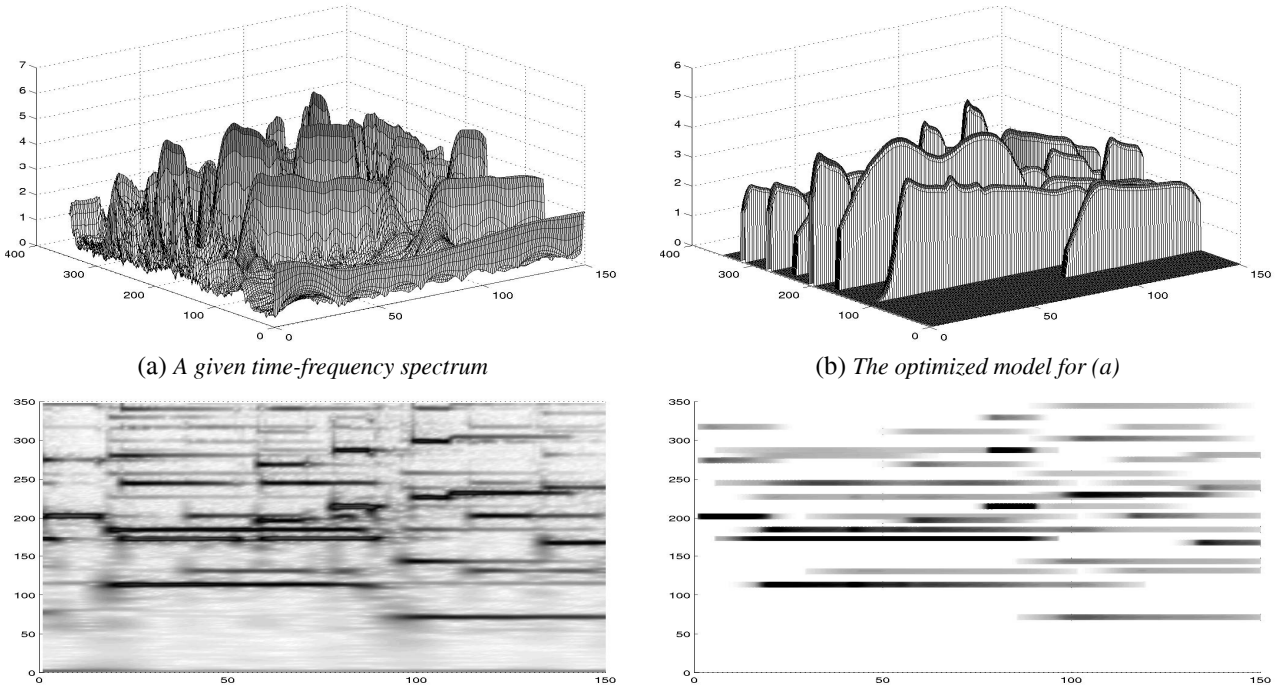
The local optimal model parameters can be effectively calculated by iteration as follows:

1. (**E-step**) Compute the auxiliary function $R(\boldsymbol{\Theta}, \boldsymbol{\Theta})$ by substituting $\widehat{\boldsymbol{\Theta}}$, the updated model parameter at the previous M-step, for $\boldsymbol{\Theta}$. This step is an update of the membership probability density $p(k, n, y | x, t, \boldsymbol{\Theta})$.

2. (**M-step**) Update the parameters to $\widehat{\boldsymbol{\Theta}}$ that maximizes the auxiliary function $R(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}})$. $\widehat{\boldsymbol{\Theta}}$ is calculated from $\partial R(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}) / \partial \widehat{\boldsymbol{\Theta}} = 0$. This step is an optimal model approximation under fixed $p(k, n, y | x, t, \boldsymbol{\Theta})$.

4. EXPERIMENTS

For the evaluation of the proposed method, it was tested on 2 pieces of real music performance data excerpted from RWC music database. Time series of power spectrum was analyzed by Gabor wavelet transform with frame shift of 20ms for input digital signals of 16kHz sampling rate. The lower bound of the frequency range and the frequency resolution were 50Hz and 16.7cent, respectively. The interval of time range of time-frequency plane was 3s(150 frames). The initial parameters of $(\mu_k, o_k | k = 1, \cdots, K)$ for EM algorithm were automatically determined by extracting 70 largest peaks from the given spectrum distribution. During the iteration of EM algorithm, the total number of audio streams were estimated by thresholding, i.e., remove every audio stream model whose weight parameter w_k becomes smaller than the threshold. A typical example of the optimized model and the corresponding time-frequency spectrum are shown in Fig.6. As seen in (b) and (d), not



(a) *A given time-frequency spectrum*

(b) *The optimized model for (a)*

(c) *A grayscale display of the given spectrogram*

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

Fig. 6. 3D and grayscale diplays of the given spectrum and the parameter-optimized model

Table 2. Results of Pitch Name Estimation

Expermiental data		Accuracy(%)	
Titles	Instruments	Previous	Proposed
Crescent Serenade	Guitar	85.3	**92.1**
For Two	Piano	79.8	**86.2**

only pitch but onset/offset time, power envelope and spectral components of every audio stream are quite appropriately estimated.

We evaluated the proposed method with a simple pitch name estimation task compared with a frame-by-frame spectrum approximation algorithm, proposed at the previous paper [5]. In this previous algorithm, each single-frame spectrum is approximated by the harmonic structure model $\sum_{k=1}^{K} h_k(x)$ using EM algorithm, and then Hidden Markov Model is applied to assemble each pitch estimates to form overall multi-pitch trajectories. Thus the comparison between the new and the previous methods may indicate the effectiveness of our new idea. Although the experiment was done with very limited test data, the results presented in table 2 show significant improvement in pitch estimation accuracy over our previous algorithm.

5. CONCLUSION

We presented a new solution to the multi-pitch analysis problem based on time-frequency clustering using Gaussian kernel geometric spectrum model. The approach addressed in this paper has just been established and has many possibilites for future extensions. Flexible modifications are possible within the same framework only by relaxing the assumptions made in this paper, e.g., (1) modeling a pitch trajectory curve with polynomial, (2) allowing a different power envelope curve among all partial component and (3) introducing inharmonicity factor, etc. Estimating the number of audio streams is also an interesting issue, which will be one of the next directions of our work.

6. REFERENCES

[1] A. Klapuri, T. Virtanen and J. Holm, "Robust Multipitch Estimation for the Analysis and Manipulation of Polyphonic Musical Signals," *In Proc. COST-G6 Conference on Digital Audio Effects*, pp. 233–236, 2000.

[2] K. Nishi, S. Ando and S. Aida, "Optimum Harmonics Tracking Filter for Auditory Scene Analysis," Proc. IEEE, ICASSP 96, pp. 573–576, 1996.

[3] S. Godsill and M. Davy, "Baysian Harmonic Models for Musical Pitch Estimation and Analysis," *Proc. ICASSP2002*, Vol. 2, pp. 1769–1772, 2002.

[4] M. Goto, "A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation Using EM Algorithm for Adaptive Tone Models," *Proc. ICASSP2001*, Vol. 5, pp. 3365–3368, 2001.

[5] H. Kameoka, T. Nishimoto and S. Sagayama, "Separation of Harmonic Structures Based on Tied Gaussian Mixture Model and Information Criterion for Concurrent Sounds," Proc. ICASSP2004, AE-P5.9, May 2004.