Statistical Harmonic Model with Relaxed Partial Envelope Constraint for Multiple Pitch Estimation

○Jun Wu, Yu Kitano, Stanisław Andrzej Raczyński, Shigeki Miyabe, Takuya Nishimoto, Nobutaka Ono and Shigeki Sagayama (The University of Tokyo)

1 Introduction

The goal of multiple pitch estimation is to estimate fundamental frequencies of multiple harmonic signals simultaneously present in an input musical signal. Some multipitch analyzers such as multiagent-based [1], cochlear filtering-based [2], parametric signal and spectrum modeling-based approaches [3], non-negative matrix factorization (NMF) [4] and harmonic temporal clustering (HTC) [5] have been proposed for practical application of multipitch estimation. The main problem with all of these methods is related to the noise-like nature of the attack phase of otherwise harmonic notes—so far it was not given any attention from researchers.

In this paper, a harmonic model capable of modeling both harmonic part and the inharmonic transient sounds occurring during the attack stage of note is proposed. The model uses a mixture of Gaussians to represent temporal structure of the notes and enables estimating expectation of every parameter by means of the EM algorithm. The paper organization is as follows. In Section 2, the proposed model is introduced. In section 3, the experimental results are demonstrated and compared with previous researches. Finally, the conclusion is given in section 4.

2 Statistical Harmonic Model

The motivation of this research is to design a model capable of modeling both harmonic and inharmonic partial from music. The inharmonic phase affects in multipitch estimation but it is not given enough attention in the previous research. The proposed model approximates the observed power spectrogram of input music signal *W(x;t)* (where *x* is log-frequency and *t* is time) with a sum of *K* parametric models, each of which represents a single continuous pitch. Every pitch model is composed of a fundamental partial (F0) and *N* harmonic partials. Let $U_{k,n}(t)$ be the time-domain envelope of the *n*th partial in the *k*th model. The frequency-domain envelope shapes of all partials are identical, but shifted by log(*n*), the frequency-domain envelope can be modeled by:

$$U_{k,n}(t)v_{k,n}\mathcal{N}(t,\tau_k+y\phi_{k,n},\sigma_k) \quad (1)$$

$$\text{Satisfying } \forall k,\sum_n v_{k,n}=1 \quad (2)$$

The time-domain envelope of *n*th partial is given by:

$$U_{k,n}(t)=\sum_y u_{k,n,y}\,\mathcal{N}(t,\tau_k+y\phi_{k,n},\sigma_k) \quad (3)$$

where τ_k is the center of the Gaussian, which is considered as an onset time estimation, $u_{k,n,y}$ is the weight parameter for each kernel, which allows the function to have variable shapes for each harmonic partial. $u_{k,n,y}$ is defined as the coefficient of the power envelope function of *k*th model, *n*th partial, *y*th kernel. The time-domain envelope of each partial is different, though there is a relationship between them. The model is expressed as a mixture of Gaussian mixture model (GMM) with constraints on the kernel distributions: supposing that there is harmonicity with *N* partials modeled in the frequency direction, and the power envelope is described using *Y* kernel distribution in the time direction. The model can be written in the form

$$q_k(x,t;\theta)=\sum_n\sum_y S_{k,n,y}(x,t;\theta) \quad (4)$$

And the Kernel distribution can be written as $S_{k,n,y}(x,t;\theta)=$

$$\frac{w_k v_{k,n}u_{k,n,y}}{2\pi\delta_k\emptyset_k}e^{-\frac{(x-\mu_k(t)-\log n)^2}{2\sigma_k^2}-\frac{(t-\tau_k-y\emptyset_{k,n})^2}{2\phi_{k,n,y}^2}} \quad (5)$$

The term *attack* is defined to represent the inharmonic phenomenon at the very beginning (onset) of notes played by harmonic instruments. During the attack phase, the harmonic structure is unclear. The onset noise model is included in the model by using the following modification:

*倍音成分の時間包絡に関する自由度を持つ統計的調波モデルによる多重ピッチ推定、呉軍、北野佑、ラチンスキ・スタニスワブ、宮部滋樹、西本卓也、小野順貴、嵯峨山茂樹(東大情報理工).

$$U'_{k,n}(t) = \frac{U_{k,n,0}}{\sqrt{2\pi}\emptyset_{k,n}} e^{-(t-y\emptyset_{k,n})^2/2\emptyset_{k,n}^2} \quad (6)$$

Therefore, the pitch is modeled as a Gaussian distribution which is correlated with the harmonic part in time domain. The attack model in frequency domain was represented by a Gaussian mixture model.

$$F(x) = \sum_{j=1}^{m} \alpha_j g(x, \mu_j, \sigma_j^2) \quad (7)$$

where $g(x, \mu_j, \sigma_j^2)$ is a component Gaussian distribution characterized by means, covariance and weight of its component distributions. We have employed the EM algorithm to estimate all of the model's parameters. We use the Kullback–Leibler (KL) divergence as the cost function;

$$J = \sum_k \iint_D m_k(x,t) W(x;t) log \frac{m_k(x,t)W(x;t)}{q_k(x,t;\theta)} \quad (8)$$

under the constraint;

$$\sum_k m_k\ (x,t) = 1, 0 < m_k\ (x,t) < 1, \forall x, \forall t \quad (9)$$

Therefore the problem is regarded as the minimization of equation (8). The M-step can be realized by the iteration of the update the parameters depending on each acoustic object, which can be obtained analytically by the combination of an undetermined Lagrange multipliers method.

3 Experimental results

To evaluate the proposed algorithm, we have performed experiments with music samples selected from the RWC music database [6]. Since the RWC database also includes the MIDI files associated with each recorded music signal data, we have synthesized audio data from the MIDI files and used the MIDI files as the ground truth, which is common practice for multipitch estimation. The pitch accuracies are computed by

$$\text{Accuracy} = \frac{\text{X}-\text{D}-\text{I}-\text{S}}{\text{X}} \quad (10)$$

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

We compared the proposed algorithm with NMF algorithm, the "PreFEst" algorithm and the original HTC. The results are shown in Table 2. The proposed algorithm outperforms the comparison algorithm for all of test data.

Table 2. Accuracy of F0 estimation algorithms

Testing data	NMF (%)	PreFEst (%)	HTC (%)	Proposed (%)
RWC-MDB-J -2001 No.9	65.4	74.2	81.2	85.4
RWC-MDB-J -2001 No.7	68.0	71.8	77.9	79.2
RWC-MDB-J -2001 No.1	60.3	55.9	64.2	68.5
RWC-MDB-J -2001 No.8	69.5	76.2	75.2	78.7
RWC-MDB-J -2001 No.2	57.5	62.3	62.2	65.3
RWC-MDB-J -2001 No.6	58.9	48.8	63.8	69.2

4 Conclusion

This paper proposed an algorithm to model the attack part in the beginning part of some notes as well as the harmonic part because the attack problem will decrease the accuracy of multipitch estimation in a lot of cases. The proposed algorithm is efficient for estimating multipitch from polyphonic music, which was proved by experiments. Applying the model to other interesting tasks is considered as future work.

References

[1] 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.

[2] 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.

[3] 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.

[4] S. A. Raczynski, N. Ono, S. Sagayama, "Multipitch analisys with Harmonic Nonnegative Matrix Approximation," Proc. of ISMIR, pp.381-386, Sep., 2007.

[5] H. Kameoka, T. Nishimoto, S. Sagayama, "A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering," IEEE Trans. on Audio, Speech and Language Processing, Vol. 15, No. 3, pp.982-994, Mar., 2007.

[6] 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.