Multipitch estimation by joint modeling of harmonic and transient sounds

Jun Wu, Emmanuel Vincent, Stanislaw Raczynski, Takuya Nishimoto, Nobutaka Ono, Shigeki Sagayama

MULTIPITCH ESTIMATION BY JOINT MODELING OF HARMONIC AND TRANSIENT SOUNDS

Jun Wu*, Emmanuel Vincent†, Stanisław Andrzej Raczyński*, Takuya Nishimoto*,
Nobutaka Ono* and Shigeki Sagayama*
*The University of Tokyo, Tokyo 113–8656, Japan
E-mail: { wu, raczynski,nishi,onono,sagayama }@hil.t.u-tokyo.ac.jp
†INRIA, Centre de Rennes - Bretagne Atlantique, 35042 Rennes Cedex, France
E-mail: emmanuel.vincent@inria.fr

ABSTRACT

Multipitch estimation techniques are widely used for music transcription and acquisition of musical data from digital signals. In this paper, we propose a flexible harmonic temporal timbre model to decompose the spectral energy of the signal in the time-frequency domain into individual pitched notes. Each note is modeled with a 2-dimensional Gaussian mixture. Unlike previous approaches, the proposed model is able to represent not only the harmonic partials but also the inharmonic attack of each note. We derive an Expectation-Maximization (EM) algorithm to estimate the parameters of this model and illustrate the higher performance of the proposed algorithm than NMF algorithm [9] and HTC algorithm [10] for the task of multipitch estimation over synthetic and real-world data.

Index Terms— multipitch estimation, GMM, EM algorithm, attack

1. INTRODUCTION

Multipitch estimation aims at estimating the fundamental frequencies and the onset times of the musical notes simultaneously present in a given music signal. It is considered to be a difficult problem mainly due to overlap of the overtones of different pitches - a common phenomenon in Western music. Numerous approaches have been tried, including perceptually motivated methods [1,2, 3,4], parametric signal model-based methods [5,6], classification-based methods [7] and parametric spectrum model-based methods [8,9,10,11]. Model-based approaches have received much interest recently due to their ability to exploit prior information about the signal structure.

While all these approaches account for the harmonic part of pitched notes, the attack part has not been given much attention in the context of multipitch estimation. This often results in multipitch estimation errors due to the fitting of inharmonic attack transients by a combination of harmonic sounds. Designing a model able to deal both with harmonic and inharmonic parts is therefore essential.

In this paper, we propose an algorithm for polyphonic pitch estimation that models both the harmonic and transient parts of musical notes with a mixture of two-dimensional, spectro-temporal Gaussians. This model is inspired by the parametric spectrum model-based algorithm in [10], which represents the power spectrum of the observed signal as a mixture of individual partial spectra. We augment this model with an attack model so as to avoid the spurious short-duration notes typically estimated at note onsets and derive an EM algorithm to estimate the time-varying fundamental frequency of each note together with the other parameters in the Maximum Likelihood (ML) sense.

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 research. Finally, the conclusion is given in section 4.

2. JOINT MODEL OF HARMONIC AND TRANSIENT SOUNDS

We assume that the input signal is represented by the power of the output of a constant-Q transform with Gabor filters. The transform is computed with a temporal resolution of 16 ms for all subbands. The lower bound of the frequency range and the frequency resolution are set to 60 Hz and 12 cents, respectively [10].

The proposed model approximates the observed nonnegative power spectrogram $W(x;t)$ (where x denotes the frequency bin and t the time frame number) with a mixture of K nonnegative parametric models, each of which represents a single musical note. Every such note model is composed of a fundamental partial (F0), N harmonic partials and an inharmonic transient. Figure 1 shows an example of power spectrogram of a piano note with the transient attack part marked with a rectangle.

The power spectrogram of the k-th note is represented by

$$q_k(x,t) = w_k \sum_n H_{k,n}(x,t) + A_k(x,t). \quad (1)$$

where w_k is the total energy of this note, $H_{k,n}(x,t)$ represents the spectrogram of the n-th harmonic partial and $A_k(x,t)$ the spectrogram of the attack part of that note. All parameters of the model are listed in Table 1.

Parameter	Physical meaning
$\mu_k\ (t)$	Pitch contour of the *k*-th note
w_k	Energy of the *k*-th note
$v_{k,n}$	Relative energy of the *n*-th partial in the *k*-th note
$u_{k,n,y}$	Power envelope coefficient of the *n*-th partial of the *k*-th note at the *y*-th time frame
τ_k	Note onset time
$Y\phi_{k,n,y}$	Duration (*Y* is constant)
σ_k	Diffusion of partials in the frequency domain
α_j	Coefficient of the *j*-th Gaussian in the *k*-th transient model

Table 1. Parameters of the proposed model.

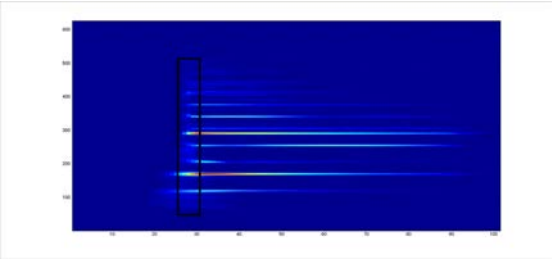

Figure 1. Example spectrogram of an isolated piano note. The attack part is emphasized by a black rectangle.

2.1. Harmonic model

The harmonic part of the proposed model is based on the one described in [10]. However, in contrast to [10], the temporal envelope can be different for each of the partials, which results in closer fit to the observed musical notes.

The harmonic model of each partial $H_{k,n}(x,t)$ is defined as the product of a spectral model $F_{k,n}(x)$ and a temporal model $U_{k,n}(t)$. Since the constant-Q Gabor transform is used as our input, the spectral model follows a Gaussian distribution centered on its log-frequency, as illustrated in Figure 2. Given the fundamental log-frequency $\mu_k\ (t)$ of *k*th note, the log-frequency of the *n*th partial is given by $\mu_k\ (t) + \log n$ (see Figure 2). This results in

$$F_{k,n}(x) = v_{k,n}\mathcal{N}(x, \mu_k(t) + \log n, \sigma_k) \quad (2)$$

where $v_{k,n}$ is the relative power of the *n*th partial satisfying

$$\forall k, \textstyle\sum_n v_{k,n} = 1 \quad (3)$$

The temporal model of each partial is designed as a Gaussian Mixture Model (GMM) with constrained expected values: the number of Gaussians is fixed to *Y* and the means are uniformly spaced over the duration of the note. This results in

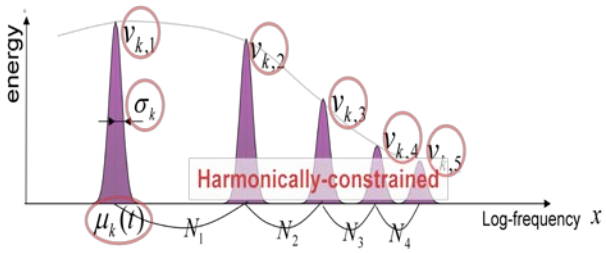

Figure 2. Cutting plane of $q_k(x,t;\theta)$ at time *t*.

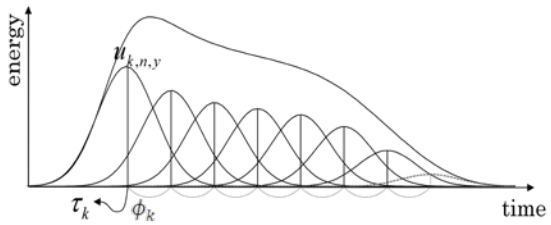

Figure 3. Power temporal envelope $U_{k,n}(t)$ at frequency *x*.

$$U_{k,n}(t) = \textstyle\sum_y u_{k,n,y}\, \mathcal{N}(t, \tau_k + y\phi_k, \sigma_k) \quad (4)$$

where τ_k is the center of the first Gaussian (considered to be the estimate of the onset time) and $u_{k,n,y}$ is the weight for each time frame that allows the temporal envelope to have a variable shape for each partial. The weight parameters are normalized to satisfy

$$\forall k, \forall y: \textstyle\sum_y u_{k,n,y}\ (x,t) = 1. \quad (5)$$

An example temporal envelope is depicted in Figure 3.

2.2. Transient model

We now define the transient model $A_k(x,t)$ as the product of a spectral model *F'(x)* which does not depend on the note number *k* but on the associated instrument only and a temporal model $U'_k(t)$, which is another Gaussian

$$U'_k(t) = \mathcal{N}(t, \tau_k, \sigma_k) \quad (6)$$

Because the inharmonic transient occurs at the same time as the onset of harmonic partials, the parameters in this distribution are constrained to be equal to the first component of the temporal harmonic model.

The spectral model is represented by a GMM

$$F'(x) = \textstyle\sum_{j=1}^{J} \alpha_j \mathcal{N}(x, \mu_j, \sigma^2) \quad (7)$$

where the weights α_j encode the spectral shape. μ_j and σ are fixed in similar fashion to the temporal harmonic model, where the spacing between successive Gaussians is equal to their standard deviations.

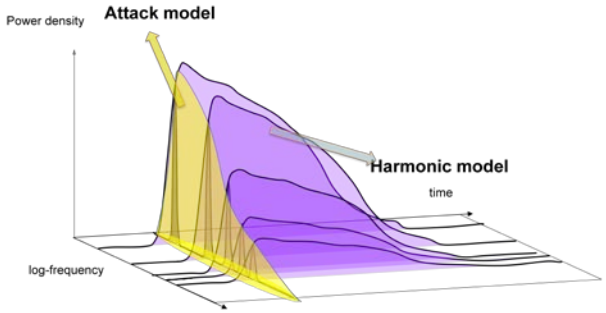

Figure 4. Representation of the proposed model.

	McGill	RWC	Uiowa	Total
bassoon	16	112	113	241
cello	40	430	337	807
clarinet	47	120	423	590
flute	90	36	226	352
oboe	27	34	104	165
piano	67	88	88	243
tuba	16	90	111	217
viola	32	467	271	770
violin	93	45	283	421

Table 2. Number of notes from each of the isolated note databases used to generate the synthetic test set.

2.3. Joint model

The harmonic model is a GMM both in the time and the frequency domain, while the transient model is a GMM in the frequency domain only. Overall, this can be expressed as

$$q_k(x,t;\theta) = \sum_n \sum_z S_{k,z}(x,t;\theta) \quad (8)$$

where z indexes $N \times Y + J$ Gaussians representing either a harmonic or a transient component, and θ denotes the full set of parameters of all notes. The entire signal can therefore be modeled by a single mixture of Gaussians $S_{k,z}(x,t;\theta)$. The resulting spectrum model is shown in Figure 4.

2.4. Inference with the EM algorithm

We employed the EM algorithm to estimate all of the model parameters. We assume that the observed energy density *W(x;t)* has an unknown fuzzy membership to the *k*th note, introduced as a spectral masking function $m_k(x,t)$. To minimize the difference between the observed power spectrogram *W(x;t)* and the note models, we employ the commonly used Kullback–Leibler (KL) divergence as the global cost function:

$$J = \sum_k \iint_D m_k(x,t) W(x;t) \log \frac{m_k(x,t) W(x;t)}{q_k(x,t;\theta)} \quad (9)$$

under the constraint that

$$\sum_k m_k\,(x,t) = 1, 0 \le m_k\,(x,t) \le 1, \forall x, \forall t. \quad (10)$$

The problem of multipitch transcription can therefore be regarded as the minimization of (9).

The E-step consists of estimating $m_k(x,t)$ while the M-step consists of iteratively updating the parameters of each note model using analytical update rules similar to those in [10]. These update rules can easily be derived by means of Lagrange multipliers but can unfortunately not be listed here due to lack of space.

3. EXPERIMENTAL EVALUATION

We evaluated the performance of the proposed algorithm for the task of multipitch estimation over both synthetic and recorded performance data. The synthetic dataset was built from three databases: the RWC Musical Instrument Sounds database [12], McGill University Master Samples [13] and the University of Iowa database [14]. A large set of isolated notes was selected from these three databases. The number of notes taken from each database is listed in Table 2. For each instrument, we generated 60 single-instrument signals containing 3 or more notes. These signals consisted either of notes with similar onset times (overlapping) or notes occurring in a sequence. The obtained single-instrument signals were subsequently randomly mixed to form multi-instrument test signals of 6-second duration. In addition, we used the recorded performance data from the development set of the Multiple Fundamental Frequency (Instrument Tracking) task of MIREX 2007 [15]. We mixed individual tracks from different instruments to form additional test signals of 6-second duration. For both synthetic and real-world database, we added together 2 single-instrument signals to obtain a multi-instrument signal. We generated 80 synthetic mixtures and 40 real-world mixtures in total. The mean number of simultaneously present notes in a given time frame is 3. The minimum number is 2 and maximum number is 5. In the proposed model, the number of source models K is initialized as 60.

Thanks to the employed dataset creation procedure, the true pitches were known and could be directly compared with the estimated pitches. A returned pair of pitch and onset time was assumed to be correct if it was within 1/4 tone and 50ms of a true note. Two evaluation metrics were calculated: recall R and precision P. The latter is a measure of how many of the detected notes were correct (it indicates the number of spurious notes), while recall is a measure of how many of the true notes were detected (it indicates the number of omitted notes). The F-measure is calculated from these two values as their harmonic mean: F = 2RP/(R + P).

We have compared the proposed model with the NMF algorithm in [9] and the original HTC algorithm from [10], which achieved the highest score in the task of Multiple Fundamental Frequency Estimation at MIREX 2009. The results are shown in Tables 3 and 4. The proposed algorithm outperformed NMF by 12.5 percent units for synthetic and by 15.2 percent units for recorded performance data. It also

	P (%)	R (%)	F (%)
NMF [9]	72.5	74.4	73.4
HTC [10]	82	78.7	80.3
Proposed	85.3	86.5	**85.9**

Table 3. Multipitch estimation performance over synthetic data.

	P (%)	R (%)	F (%)
NMF [9]	44.1	46.6	45.3
HTC [10]	57.4	51.3	54.2
Proposed	59.7	61.4	**60.5**

Table 4. Multipitch estimation performance over real-world data.

outperformed HTC by 5.6 percent units for synthetic and 6.3 percent units for recorded performance data.

It is worth noting that both the obtained recall and precision were better and that the recall of the proposed algorithm improved both for synthetic data and real-world data compared to the original HTC algorithm. This is mainly due to the removal of spurious short-duration notes erroneously estimated at note onsets.

4. CONCLUSION

We have proposed a model based on the clustering principle using a harmonically structured Gaussian mixture model. The expected values of internal parameters directly correspond to such qualities as the pitch and the model is used to explain the observed short-time power spectrogram. The proposed algorithm models the harmonic part of notes, but also included a model of the initial inharmonic transients, which are prone to decrease the F0 estimation accuracy. The proposed algorithm is intuitive and the obtained results suggest that it is also efficient in estimating multiple pitches from polyphonic musical signals. The model can be used not only for musical signals, but also speech or other common sound signals. We plan to apply our model to other interesting tasks in the future.

5. ACKNOWLEDGEMENT

This work was supported by INRIA under the Associate Team Program VERSAMUS (http://versamus.inria.fr/).

6. REFERENCES

[1] W. M. Hartmann, "Pitch, periodicity, and auditory organization," *Journal of the Acousical Society of America*, vol. 100, no. 6, pp. 3491–3502, 1996.

[2] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 11, no. 6, pp. 804-816, November 2003.

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

[4] T. Tolonen, M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, 2000.

[5] M. Davy, S. J. Godsill, and J. Idier, "Bayesian analysis of western tonal music," *Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2498–2517, 2006.

[6] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 728–731, 1993.

[7] G. E. Poliner and D. P. W. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP Journal on Applied Signal Processing*, vol. 2007, article ID 48317, 2007.

[8] 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.

[9] S. A. Raczyński, N. Ono, S. Sagayama, "Multipitch analysis with Harmonic Nonnegative Matrix Approximation," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, pp.381-386, Sep., 2007.

[10] H. Kameoka, T. Nishimoto, S. Sagayama, "A multipitch analyzer based on Harmonic Temporal Structured Clustering," *IEEE Transactions on Audio, Speech and Language Processing*, vol.15, no.3, pp. 982–994, Mar, 2007.

[11] E. Vincent, N. Bertin and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 528-537, 2010.

[12] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music database," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, pp. 287–288, 2002.

[13] http://www.music.mcgill.ca/resources/mums/html/MUMS_audio.htm

[14] http://theremin.music.uiowa.edu/MIS.html.

[15] http://www.music-ir.org/mirex/wiki/2007:Multiple_Fundamental_Frequency_Estimation_%26_Tracking