MULTIPLE FUNDAMENTAL FREQUENCY ESTIMATION OF POLYPHONIC MUSIC SIGNALS

Chunghsin YEH, Axel RÖBEL & Xavier RODET

IRCAM
Analysis-Synthesis team
1, Place Igor-Stravinsky 75004 Paris France

ABSTRACT

This article is concerned with the estimation of fundamental frequencies, or *F0*s, in polyphonic music. We propose a new method for jointly evaluating multiple *F0* hypotheses based on three physical principles: harmonicity, spectral smoothness and synchronous amplitude evolution within a single source. Based on the generative quasiharmonic model, a set of hypothetical partial sequences is derived and an optimal assignment of the observed peaks to the hypothetical sources and noise is performed. Hypothetical partial sequences are then evaluated by a score function which formulates the guiding principles in a mathematical manner. The algorithm has been tested on a large collection of artificially mixed polyphonic samples and the results show the competitive performance of the proposed method.

1. INTRODUCTION

Automatic transcription of polyphonic music has attracted involvements in several research topics including multiple fundamental frequency estimation, onset detection, rhythm/meter estimation, etc.. Despite increasing research activities with respect to polyphonic music signals, the estimation of multiple *F0*s remains a challenging problem. Some of the generally admitted difficulties are: estimating the number of *F0*s, retrieving reliable time-frequency properties and treating mixtures of transient parts and stationary parts. These difficulties mainly come from the multi-timbre mixture of various musical instruments, diverse spectral characteristics which are related to different playing techniques, chords consisting of harmonically related *F0*s and acoustical interference such as reverberation.

In the present investigation we focus on estimating *F0s* in musical signals when the number of *F0*s is known in advance. The importance of including higher level features in addition to periodicity/harmonicity in multiple *F0* estimation has been demonstrated by most of the existing approaches. Martin has proposed a blackboard system gathering all available knowledge to rate *F0* hypotheses [1]. Goto [2] introduces tone models as a constraint on relative partial amplitudes. In Klapuri's multiple *F0* estimation algorithm, the spectral smoothness principle is a key to deal with overlapped partials [3]. For the probabilistic signal modeling approach presented in [4], the prior distributions of the model parameters are in fact physical constraints on spectral models in search. The core of our multiple *F0* estimation is a score function which jointly evaluates multiple *F0* hypotheses. Based on a generative quasiharmonic spectral model, hypothetical partial sequences are constructed and evaluated based on three physical principles: harmonicity, spectral smoothness and synchronous amplitude evolution within a single source.

This paper is organized as follows. In section **2**, the generative quasiharmonic model is described and the principles for *F0* estimation are established. In section **3**, we introduce a frame-based *F0* estimation method using a physical principle driven score function. In section **4**, experimental results are shown, which proves the competitive performance of the proposed method. Finally, we discuss this method and draw conclusions.

2. GENERATIVE QUASIHARMONIC MODEL

The proposed algorithm is based on a polyphonic quasiharmonic signal model of the following form

$$y[n] = \Big\{ \sum_{m=1}^{M} \sum_{h_m=1}^{H_m} a_{m,h_m}[n] \cos\big((1+\delta_{m,h_m}) h_m \omega_m n + \phi_m[n]\big)\Big\} + v[n], \qquad (1)$$

where n is the discrete time index, M is the number of sources, H_m is the number of partials for the m-th source, ω_m represents the *F0* of source m and $\phi_m[n]$ denotes the phase. The score function used here makes use of $a_{m,h}[n]$ and $\delta_{m,h}$ which are the time varying amplitude and the constant frequency detuning of the h-th partial, and also $v[n]$ which is the residual noise component. Generally it is supposed that the noise is sufficiently small such that a considerable part of the individual sinusoidal components can be identified.

Similar to [5] we understand the observed spectrum as generated by sinusoidal components and noise, and all necessary information for *F0* estimation is to be extracted from the properties of spectral peaks. Each peak is considered either sinusoidal or noise. A sinusoidal peak is assigned to one or more of the M sources in eq.(1), all unassigned peaks contribute to the noise component $v[n]$. Based on this model and given the observed spectrum and M, the most plausible *F0* hypotheses are going to be inferred. To construct and evaluate hypothetical sources, we rely on the three physical principles:

Principle 1: *Spectral match with low inharmonicity*. For each *F0* hypothesis, a hypothetical partial sequence, *HPS*, is constructed by selecting harmonically matched peaks from the observed spectrum in such a way that $\delta_{m,h}$ are minimized. The set $\{HPS_{F0_m}\}_{m=1}^{M}$ corresponding to M hypothetical sources should combinatorially explain as many peaks as possible of the observed spectrum such that the remaining noise energy is minimized.

0-7803-8874-7/05/$20.00 ©2005 IEEE

Principle 2: *Spectral smoothness*. The spectral envelopes of musical instrument sounds tend to form smooth contours [3]. While constructing the *HPS* of a source, the partials should be selected in a way such that $\{a_{m,h}\}_{h=1}^{H_m}$ results in a spectral envelope as smooth as possible.

Principle 3: *Synchronous amplitude evolution within a single source*. Partials belonging to the same source should have similar time evolution of the amplitudes $\{a_{m,h}\}_{h=1}^{H_m}$ contained in a *HPS*. If the partials assigned to a hypothetical source match mostly to noisy peaks, they evolve in a random manner and thus will not have a synchronous amplitude evolution.

3. MULTIPLE *F0* ESTIMATION

Based on the three principles described above, we design a multiple *F0* estimation system. The main task is to formulate these principles into four criteria serving as the core components in a score function for evaluating the plausibility of one set of *F0* hypotheses.

3.1. Front end

While analyzing polyphonic signals with limited spectral resolution, one often observes that the dense distribution of partials causes some peaks to be hidden by relatively larger coincident ones. Thus, we evaluate the shapes of the observed peaks and their spectral properties proposed in [6] to choose the possibly overlapped peaks which are then processed to extract hidden peaks.

To generate an *F0* candidate list, we use a harmonic matching technique. For each *F0* hypothesis, a vector d_{F0} is constructed to evaluate the degree of deviation from a harmonic model to the observed peaks, and a tolerance interval around each harmonic is used to measure the goodness of harmonic matching. For the i-th observed peak matching the h-th harmonic, the degree of deviation is formulated as

$$d_{F0}(i) = \frac{|f_{peak}(i) - f_{model}(h)|}{\alpha \cdot f_{model}(h)} \tag{2}$$

where $f_{peak}(i)$ is the frequency of the i-th observed peak, $f_{model}(h)$ is the frequency of the h-th harmonic of the model, and α determines the tolerance interval. If an observed peak situates outside the corresponding tolerance interval, it is regarded as unmatched and $d_{F0}(i)$ is set to 1. Then we define the harmonic matching function as:

$$HAR_{F0} = \sum_{i=1}^{I} \frac{Corr(i) \cdot Spec(i) \cdot d_{F0}(i)}{\sum_i [Corr(i) \cdot Spec(i)]} \tag{3}$$

where I is the number of observed peaks, $Corr$ is the complex correlation between each observed peak and an ideal sinusoidal peak defined by the analysis window, $Spec$ is the peak energy vector. Since inharmonicity exists in most of the string instruments, it is necessary to dynamically adapt the frequencies of model harmonics according to the best matched peak. This is realized by the partial selection technique. We start with the fundamental by simply assigning it to the closest peak observed. For the following partials we consider two candidate peaks: the one closest to $f_{model}(h)$ and the one of which the mainlobe contains $f_{model}(h)$. Compared to the formerly selected partials, the peak candidate forming a smoother envelope is selected as the best matched peak. Then $f_{model}(h+1)$ is calculated by means of adding *F0* to the frequency of the best matched peak. If there is no peak assigned to the current partial, $f_{model}(h) + F0$ is used for the next match. All the peaks having been assigned to an *F0* hypothesis are forming the *HPS*. The *F0* hypotheses corresponding to local maxima of the harmonic matching function are added to the candidate list. Then, all possible combinations of the candidates will be evaluated by a score function.

Although the *HPS* of each *F0* candidate has been constructed during harmonic matching, the overlapped partials need to be taken care of. The treatment of overlapped partials is based on the idea that an overlapped partial still carries important information for at least the *HPS* that locally has the strongest energy. Therefore, the algorithm aims to assign the overlapped partial to this *HPS*. Constructing a *HPS* in fact utilizes *Principle 2* and the knowledge of spectral locations where partial overlaps may occur according to the current set of *F0* hypotheses under investigation. The guiding principle is to make use of the credibility of available information. The strategy for treating the overlapped partials is listed below:

(i) Partials having potential collision are determined from each hypothetical combination of *HPS*s.

(ii) The local energy strength of the envelope is obtained by means of interpolating the neighboring partial amplitudes that are not collided. By comparing the interpolated amplitudes estimated from all the *HPS*s of a hypothetical set of *F0* candidates, the overlapped partial is exclusively assigned to the one having the most dominant interpolated amplitude among all and then labeled as "effective" which means that it can be used for interpolation for its neighboring partials. The rest of the *HPS*s the overlapped partial is considered "not effective" and is labeled as existing but without a specified partial amplitude.

(iii) If one neighboring partial happens to be overlapped and not effective, the non-overlapped partial at the other side is used instead. If the two neighboring partials are not effective, the corresponding *HPS* is not considered as having reliable information for interpolation and is thus excluded.

(iv) If the amplitude of the overlapped partial is smaller than any interpolated amplitude, it is difficult to infer which *F0* hypothesis contributes the most and thus partial assignment is not carried out but this overlapped peak in all *HPS*s are labeled as "effective" for further use of interpolation.

3.2. The score function

Having constructed the most reasonable *HPS*s for each set of *F0* hypotheses, we design a score function to rank these hypothetical sets. The score function formulates the three principles into four criteria: harmonicity *HAR*, mean bandwidth *MBW* and effective length *EFL* of *HPS*s, and the standard deviation of mean time *SYNC*.

HAR is an indication of harmonicity and totally "explained" energy. It is formulated as eq.(3) with $d_{F0}(i)$ replaced by

$$d_M(i) = min\big(\{d_{F0_m}(i)\}_{m=1}^{M}\big) \tag{4}$$

That is, each observed peak is matched with the closest partial among those of $\{HPS_{F0_m}\}_{m=1}^{M}$ and thus each combination under investigation can perform its optimal match.

To evaluate the smoothness of a *HPS*, we use the mean bandwidth as a criterion. Each *HPS* is assembled with its flipped sequence to construct S_{F0_m} for further evaluation. Applying K-point FFT to S_{F0_m}, we obtain the linear spectrum X_{F0_m} and calculate the mean bandwidth MBW_{F0_m} as

$$\mathit{MBW}_{F0_m} = \sqrt{2 \cdot \frac{\sum_{k=1}^{K/2} k[X_{F0_m}(k)]^2}{\sum_{k=1}^{K/2} [X_{F0_m}(k)]^2}} \qquad (5)$$

This indicates the degree of energy concentration in the low frequency region and thus S_{F0_m} with less variation results in a smaller value of MBW_{F0_m}.

For the signal produced by a musical instrument, the spectral centroid tends to lie around lower partials because higher partials often decay gradually. From this general principle related to *Principle 2*, we can similarly evaluate the energy spread of a *HPS* in terms of the effective length of S_{F0_m}. Instead of removing the non-reliable components from HPS_{F0_m}, we use linear interpolation to reconstruct an estimated partial sequence EPS_{F0_m}. Then the effective length of HPS_{F0_m} can be calculated as

$$\mathit{EFL}_{F0_m} = \sqrt{2 \cdot \frac{\sum_{n=1}^{N_m} n[\mathit{EPS}_{F0_m}(n)]^2}{L \cdot \sum_{n=1}^{N_m} [\mathit{EPS}_{F0_m}(n)]^2}} \qquad (6)$$

where N_m is the length of EPS_{F0_m}. L is a normalization factor determined by $\lfloor F_{90}/F0_{min} \rfloor$, where F_{90} stands for the frequency limit containing 90% of spectral energy in the analyzing frequency range and $F0_{min}$ is the minimal hypothetical *F0* in search. Since the spectral envelopes of musical signals are not always smooth, this criterion functions as the further test of physical consistency of *Principle 2* and acts as a penalty function for subharmonics which explain more than one source in the observed spectrum.

To evaluate the synchronicity of the temporal evolution of the hypothetical sinusoidal components in a *HPS*, we rely on the estimation of the mean time for individual spectral peaks. Mean time is an indication of the center of gravity of signal energy [7] and the mean time of a spectral peak can be used to characterize the amplitude evolution of the related signal [8]. For a coherent *HPS* we expect synchronous evolution resulting in a small variance of mean time concerning the collection of peaks. The mean time of a hypothetical source, denoted as T_{F0_m}, is calculated as the power spectrum weighted sum of the mean time of the hypothetical partials. The standard deviation of mean time of the partials in HPS_{F0_m} is then formulated as

$$\mathit{SYNC}_{F0_m} = \frac{1}{win/2} \sqrt{\sum_{i=1}^{I} \{[\bar{t}_i - T_{F0_m}]^2 \cdot w_{F0_m}(i)\}} \qquad (7)$$

where win is the window size, $\bar{t}_i$ denotes the mean time of the i-th observed peak. The weighting vector $\{w_{F0_m}(i)\}_{i=1}^{I}$, normalized to be summed to one, is constructed from HPS_{F0_m} by setting zeros for the following components: (i) non-reliable partials due to overlaps and (ii) close partials of which spectral phases are probably disturbed. Lastly, $\{w_{F0_m}(i)\}_{i=1}^{I}$ is compressed by an exponential factor to reduce the dynamic range such that the significance of spurious peaks is raised. This makes use of the spurious peaks to penalize more a *HPS* containing more spurious peaks.

Here we define "effective energy", denoted as Eengy_{F0_m}, for each *F0* hypothesis as the sum of linear amplitudes of HPS_{F0_m}. Then $\{\mathit{MBW}_{F0_m}\}_{m=1}^{M}$, $\{\mathit{EFL}_{F0_m}\}_{m=1}^{M}$ and $\{\mathit{SYNC}_{F0_m}\}_{m=1}^{M}$ of a set of *F0* hypotheses are weighted by the effective energy and then summed to define *MBW*, *EFL* and *SYNC*, respectively. The final score function is formulated as

$$D = p_1 \cdot \mathit{HAR} + p_2 \cdot \mathit{MBW} + p_3 \cdot \mathit{EFL} + p_4 \cdot \mathit{SYNC} \qquad (8)$$

where $\{p_j\}_{j=1}^{4}$ are the weighting parameters for the four criteria. These criteria are designed in a way that a smaller weighted sum stands for higher score. Notice that *HAR* favors lower hypothetical *F0*s while *MBW*, *EFL* and *SYNC* favor higher ones. Therefore, the criteria perform in a complementary way and the weighting parameters should be optimized to balance the relative contribution of each criterion such that the score function generally supports the correct combinations of *F0*s the best. Similar to the *F0* refining technique in [5], we apply a linear regression of *F0*s estimated from the effective hypothetical partials.

4. EXPERIMENTAL RESULTS

To evaluate the proposed *F0* estimation method, we perform a frame-based test using mixtures of musical samples. Non-transient parts of monophonic musical samples are pre-selected and then mixed with equal mean-square energy to generate polyphonic samples. Estimation of a polyphonic sample is performed within a single frame. The number of *F0*s is given in advance for the *F0* estimation system to find the most probable set of *F0* hypotheses.

The parameters to be optimized are the weighting parameters $\{p_j\}_{j=1}^{4}$ in the score function and α for determining the tolerance interval in eq(2). 300 polyphonic samples containing 100 samples for each voice mixture are generated by randomly mixing musical instrument samples from the University of Iowa[1]. Then the parameters are optimized utilizing the evolutionary algorithm [9] and the set of parameters of the best performance($\{p_j\}_{j=1}^{4} = \{0.3774, 0.2075, 0.2075, 0.2075\}, \alpha = 0.035$) is used for the final evaluation on a large database. Specifications for this test are described below:

- Three databases: two-voice, three-voice and four-voice mixtures, labeled as TWO, THREE and FOUR respectively, are generated using musical samples from McGill University[2], Iowa University and IRCAM (Studio On Line). In combining M-voice polyphonic samples, M out of twelve tone names are preliminarily assigned and the monophonic samples ranging from 65Hz to 2000Hz are randomly mixed. Totally around 4800 samples are generated in a way that each combination of tone names are of equal proportion. Musical instrument samples not fitting the quasiharmonic model are excluded, such as the mallet percussion instruments and the bells [10]. To facilitate comparison, the database is published on the author's web page [11].
- The search range for *F0* is set from 50Hz to 2000Hz and the observed spectrum is analyzed up to 5000Hz. A Blackman window is used for analysis.
- *F0* reference values are created from single *F0* estimation of monophonic samples before mixing. A correct estimate should not deviate from the corresponding reference value by more than 3%. The error rates are computed as the number of wrong estimates divided by the total number of target *F0*s.

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

[2] `http://www.music.mcgill.ca/resources/mums/html/`

The testing results using two analysis window sizes, 186ms and 93ms, are shown in Table 1. Since musical samples mixed randomly surely contain harmonically related notes, we present the error rates for two groups of samples: one group of mixtures containing harmonically related notes, labeled as "harmonical", and another group "non-harmonical". The overall error rates are shown in the "total" columns. The percentages of samples in the group "harmonical" are 21.89%, 51.62% and 55.54% for the three databases TWO, THREE and FOUR, respectively.

polyphony	window	non-harmonical	harmonical	total
TWO	186ms	0.38%	2.90%	0.93%
	93ms	1.19%	4.05%	1.82%
THREE	186ms	1.03%	5.11%	3.13%
	93ms	3.31%	6.78%	5.10%
FOUR	186ms	1.78%	6.61%	4.46%
	93ms	5.77%	11.41%	8.90%

Table 1. *F0* estimation results

The errors in the group "non-harmonical" are quite small which proves the competitive performance of the proposed method. The result also demonstrates the possibility of estimating harmonically related *F0*s for the case when the mixing notes are of similar energy. To study the significance of each criterion except *HAR* in the score function, we perform three further tests by deactivating one of *MBW*, *EFL* and *SYNC* in each test. The comparison with the original result is shown in Figure 1. It is observed that the deactivation of any of the three criteria degrades the overall performance.

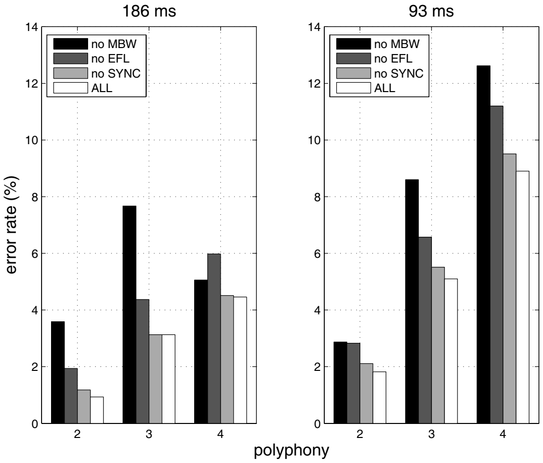

Fig. 1. Performance comparison while each criterion is deactivated

A great proportion of errors are caused by the ambiguity concerning target *F0*s and their subharmonics or superharmonics. Polyphonic samples mixed with musical instrument samples of rich resonances do not match the generative quasiharmonic model and thus are difficult to evaluate. If there exist strong resonances in addition to the partials (as observed in the string instruments), it is difficult to distinguish which part of energy relates to resonances while evaluating *HAR*. If strong resonances boost certain partials too much (as observed in oboe) and thus introduce more variations in the spectral envelope, the combination of *MBW* and *EFL* might still risk to reduce the score too much. If the partials are far less dominant than the fundamental (as generated by plucking string instruments), it is more likely that the properties of the weak partials are more noise-like and *SYNC* does not present a fair indication of the synchronous evaluation of partial amplitudes.

5. CONCLUSION

We have presented a new method to estimate multiple *F0*s for musical signals based on three physical principles. The three principles could be interpreted as reasonable prior distributions for all parameters in the generative spectral model. Instead of using an analytical approach, we optimize each hypothetical partial sequence based on these principles and then compare the credibility of possible combinations among *F0* hypotheses using a score function. Evaluation over a large polyphonic database has shown encouraging results. In order to complete the *F0* estimation system, we are continuing our studies of the estimation of the number of sources and the integration of temporal information.

6. REFERENCES

[1] Keith D. Martin, "Automatic transcription of simple polyphonic music: robust front end processing," *MIT Media Laboratory Perceptual Computing Section Technical Report*, , no. 399, 1996.

[2] Masataka Goto, "A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation using EM Algorithm for Adaptive Tone Models," in *Proc. IEEE-ICASSP 2001*, Salt Lake City, Utah, 2001, pp. V–3365–3368.

[3] Anssi P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 804–816, 2003.

[4] M. Davy and S. Godsill, "Bayesian Harmonic Models for Musical Signal Analysis," in *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, Valencia, Spain, 2003.

[5] Boris Doval and Xavier Rodet, "Estimation of fundamental frequency of musical sound signals," in *Proc. IEEE-ICASSP 91*, Toronto, 1991, pp. 3657–3660.

[6] A. Röbel and M. Zivanovic, "Signal decomposition by means of classification of spectral peaks," in *Proc. of the International Computer Music Conference (ICMC'04)*, Miami, Florida, 2004.

[7] Leon Cohen, *Time-frequency analysis*, Prentice Hall, 1995.

[8] Axel Röbel, "A new approach to transient processing in the phase vocoder," in *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx'03)*, London, 2003, pp. 344–349.

[9] Hans-Paul Schwefel, *Evolution and Optimum Seeking*, Wiley & Sons, New York, 1995.

[10] N. F. Fletcher and T. D. Rossing, *The physics of musical instruments*, Springer-Verlag, New York, 2nd. edition, 1998.

[11] `http://recherche.ircam.fr/equipes/analyse-synthese/cyeh/demo.html.`