# A NEW SCORE FUNCTION FOR JOINT EVALUATION OF MULTIPLE *F0* HYPOTHESES

*Chunghsin Yeh*

IRCAM, Analysis-Synthesis team, France
cyeh@ircam.fr

*Axel Röbel*

IRCAM, Analysis-Synthesis team, France
roebel@ircam.fr

**ABSTRACT**

This article is concerned with the estimation of the fundamental frequencies of the quasiharmonic sources in polyphonic signals for the case that the number of sources is known. 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. Given the observed spectrum a set of *F0* candidates is listed and for any hypothetical combination among the candidates the corresponding hypothetical partial sequences are derived. Hypothetical partial sequences are then evaluated using a score function formulating the guiding principles in mathematical forms. The algorithm has been tested on a large collection of artificially mixed polyphonic samples and the encouraging results demonstrate the competitive performance of the proposed method.

## 1. INTRODUCTION

The estimation of the fundamental frequency, or *F0*, of a sound source from a given signal is an essential step for many signal processing applications. For the monophonic case there exist many approaches that achieve very high performance. Despite increasing research activities with respect to polyphonic 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, treating mixtures of transient parts and stationary parts. In the following article, we propose a new method for multiple *F0* estimation under the assumption that the number of *F0*s is known in advance.

There exist several approaches for multiple *F0* estimation. A probabilistic signal modeling approach proposed in [1] applies specific prior distributions on the model parameters, such as the frequency and the amplitude of each partial, the number of partials, the detuning factor for each sinusoidal component, etc. This approach is computationally expensive and limited results are reported. In [2], a robust multipitch estimation is achieved by means of selecting reliable frequency channels as well as reliable peaks in the normalized correlograms. This technique has been reported to work for two-voice speech and the authors conclude that the proposed algorithm could be extended to more than two pitches. Klapuri's iterative multiple *F0* estimation algorithm handles most of the difficulties like estimating the number of *F0*s and treating the overlaps of coincident partials. Promising results are reported by evaluating a variety of polyphonic musical signals.

An iterative estimation and cancellation model has been proposed by de Cheveigné earlier in [3]. He compared an iterative approach and a full search approach which performs a joint evaluation. Based on this early study and later work in [4], he reported that a joint cancellation performs better than an iterative cancellation in that a single *F0* estimation failure may lead to successive errors in an iterative estimation cancellation manner. In fact, a joint evaluation strategy provides more flexibility in solving this problem. For each set of multiple *F0* hypotheses, spectral components in the interleaved spectrum could be reasonably allocated to each *F0* hypothesis and disturbed information provided by overlapped partials could be identified and taken care of in a more accurate way.

Therefore, we propose a new method for the joint evaluation of multiple *F0* hypotheses. Based on a generative quasiharmonic spectral model, hypothetical partial sequences are constructed and evaluated using three physical principles: harmonicity, spectral smoothness and synchronous amplitude evolution within a single source. Harmonicity is the essential principle in nearly all *F0* estimation techniques. It is known that using only harmonicity, however, often causes subharmonic/superharmonic ambiguity and thus more cues are necessary to improve the estimation performance. Both Kashino [5] and Goto [6] introduce tone models as a constraint on relative partial amplitudes. Klapuri has utilized the spectral smoothness principle [7] which assumes that the spectral envelopes of natural quasiharmonic sounds are in general rather smooth. Besides the two principles applied by the above authors, we include the synchronous evolution of sinusoidal amplitudes as another principle and finally formulate these principles into a new score function to rank all hypothetical combinations, which is one important contribution of this article. The second contribution is a new proposition to make use of the hypothetical *F0*s to determine reliable information in the observed spectrum.

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 the proposed score function. In section **4**, experimental results are shown, which proves the competitive performance of the proposed method. Finally, further improvements are discussed and conclusions are drawn.

## 2. GENERATIVE QUASIHARMONIC MODEL

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

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

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, $\omega_m$ represents the *F0* of source $m$, and $\phi_m[n]$ denotes the phase. In the current context those parameters are either fixed or of minor interest. The

score function will make use of $a_{m,h_m}[n]$ and $\delta_{m,h_m}$, which are the time varying amplitude and the constant frequency detuning of the $h_m$-th partial and $v[n]$, 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 [8] we understand the observed spectrum as generated by sinusoidal components and noise. Each spectral peak is characterized by its amplitude and frequency. A sinusoidal peak is assigned to one or more of the $M$ sources in eq.(1), all unassigned peaks contribute to the noise component $v[n]$. The model supposes quasi-stationary frequency and, therefore, the sinusoidality of an observed peak is used to rate the requirement to include it into the quasiharmonic parts of the source model. Based on this model and given the observed spectrum and $M$, the most plausible *F0* hypotheses are going to be inferred. The procedure is close to the Bayesian model specified in [1], however, to prevent the huge computational requirements of numerically maximizing the likelihood a more pragmatic approach is proposed.

To construct and evaluate hypothetical sources, we use three physical principles for quasiharmonic sounds stated in the following.

**Principle 1**: *Spectral match with low inharmonicity*. For a *F0* hypothesis, a hypothetical partial sequence $HPS_{F0}$ is constructed by selecting harmonically matched peaks from the observed spectrum in such a way that $\delta_{m,h}$ are minimized. The set $\{HPS_{F0_m}\}_{m=1}^{M}$ should combinatorially "explain" the sinusoidal components in the observed spectrum. Under the assumption that the noise energy is small it is reasonable to favor *F0* hypotheses that explain more components of the observed spectrum as long as they are not contradicted by the following two principles.

**Principle 2**: *Spectral smoothness*. For natural quasiharmonic sounds, the spectral envelopes usually form smooth contours. While constructing $HPS_{F0}$ of a source, the partials should be selected in a way that $\{a_{m,h_m}\}_{h_m=1}^{H_m}$ results in a smooth spectral envelope. For partial sequences fitting well to **Principle 1**, those with smoother spectral envelopes are more probable to be originated from natural sources such as musical instruments.

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

## 3. MULTIPLE *F0* ESTIMATION

Based on the three principles described above, we design a frame-based 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 multiple *F0* hypotheses.

### 3.1. Front end

#### *3.1.1. Extracting hidden partials*

When analyzing polyphonic signals with limited spectral resolutions, one often observes that the dense distribution of partials causes some peaks be hidden by relatively larger coincident ones. Thus, extracting hidden partials is essential to increase spectral resolution, which leads to a more accurate harmonic matching in the later stage. As shown in the top of Figure 1, a peak of unsymmetric form might correspond to overlapped partials.

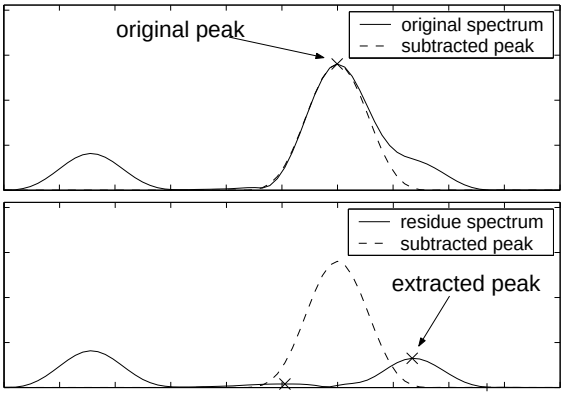


Figure 1: *Extracting the hidden partial*

To search for these hidden partials, we use a simple symmetry test for the shapes of the observed peaks. For each peak, we locate its neighboring valleys and choose the closer one to define a reference range (the bin number from one observed peak to its nearest valley). The degree of symmetry is defined as the summation of amplitude differences between the two sides of a spectral peak, considering the frequency bins within the reference range. Then a threshold is set for the degree of symmetry to select relatively unsymmetric peaks for further processing. After estimating the frequency and the frequency slope of each selected peak [9], we subtract it using the least square error criterion to extract the hidden peak as indicated in the bottom plot of Figure 1. To prevent the addition of simple residual energy as a new sinusoid, a resolved peak is kept as a successfully extracted partial only if it is not weaker than the original peak by 40 dB and should be located further than half the mainlobe width away from the original peak.

#### *3.1.2. Generating the candidate list*

To generate a *F0* hypothesis list, we use an harmonic matching technique since harmonicity is the primary concern in *F0* estimation. The harmonic matching technique matches the regular spacing between adjacent partials to determine a coherent *F0* and has been widely used for *F0* estimation in the spectral domain [10].

Given a *F0*, we construct a vector $d_{F0}$ evaluating the degree of deviation from a harmonic model to the observed peaks. A tolerance interval around each harmonic is used to measure the goodness of the harmonic match. For the $i$-th observed peak matching the $h$-th harmonic, the degree of deviation is formulated as

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

where $f_{peak}(i)$ is the frequency of the $i$th observed peak, $f_{model}(i)$ is the frequency of the $h$th harmonic of the model, and $\alpha$ determines the tolerance interval $2 \cdot \alpha \cdot f_{model}(h)$. If an observed peak situates outside the corresponding tolerance interval, it is regarded as unmatched and $d_{F0}(i)$ is set to 1.

Since inharmonicity exists in most of the string instruments, it is necessary to dynamically adapt the frequencies of model harmonics according to the matched peaks. Thus, $f_{model}(h)$ is calculated by means of adding *F0* to the previously matched peak

frequency. If not a single peak is matched for the previous partial, $f_{model}(h-1)+F0$ is used for the current match. The technique of selecting one single matched peak (among all the peaks situating in the tolerance interval) as a reference position makes use of **Principle 2** and is described later.

Three vectors are chosen to weight $d_{F0}$: (i) the complex correlation between each observed peak and an ideal peak defined by the analysis window, (ii) the linear amplitudes of the observed peaks, and (iii) an attenuation vector favoring the first several partials[1], as indicated in the top plot of Figure 2.

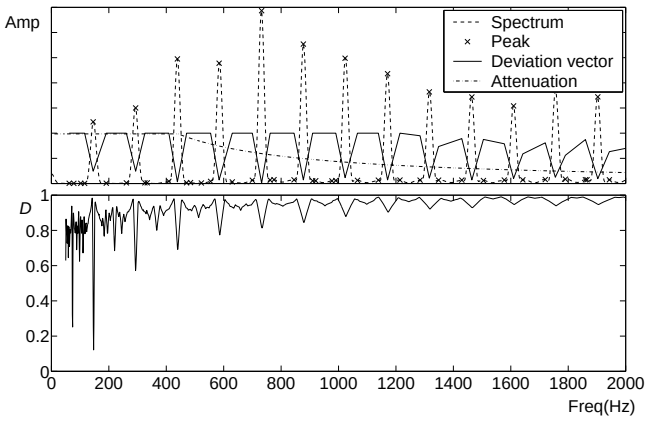


Figure 2: *Harmonic matching: a tenor trombone note at 137Hz*

The complex correlation favors peaks of better sinusoidality (shape and phase). The linear peak amplitude adjusts relative significance by considering peaks of larger energy more important. The third weighting vector attenuates less reliable matches for higher partials because they tend to be inharmonic and non-stationary. Besides, the gradual decay nature of higher partials reduces the reliability in the presence of stronger partials from other sources. Then the weighted deviation vector is summed and normalized between 0 and 1. The resulting indicator for harmonic matching is denoted as $D$. An example is shown in the bottom plot of Figure 2, the weighted sums of the deviation vectors for *F0* hypotheses ranging from 50Hz to 2000Hz are plotted. A lower value means a better match and thus higher harmonicity. The harmonic matching indicator is applied to polyphonic spectra to select *F0* candidates corresponding to local minima of $D$ for the joint evaluation.

Assume there are $P$ *F0*s in the candidate list and there are $M$ *F0*s to be estimated from the observed spectrum which results in the need to evaluate $C_M^P$ combinations of *F0* hypotheses.

#### *3.1.3. Generating Hypothetical Partial Sequences*

Constructing *HPS*s of *F0* hypotheses in the candidate list is realized by the partial selection technique. Both Parsons [11] and Duifhuis [12] have proposed selecting the nearest peak around a harmonic. However, this technique might fail if a partial is surrounded by spurious peaks and partials of other sources. Therefore, we try to increase the robustness by means of utilizing **Principle 2** and the knowledge of spectral locations where partial overlaps may occur according to the current *F0* hypotheses under investigation. The goal is to make the best of the available credible information.

[1]The third partial is tested to be a good starting point for attenuation.

The construction procedure has two steps: (i) Each *HPS* is constructed by assigning the most plausible peaks, and (ii) the overlapped partials containing less credible amplitudes are removed from *HPS* to ensure reliability for evaluating the spectral envelope in the score function.

To construct a *HPS* we start with the first partial by simply assigning it to the closest peak observed. For the following partials we consider two candidate peaks: the closest one and the one of which the mainlobe contains the corresponding harmonic position. Compared to the formerly selected partials, the peak candidate forming a smoother envelope is sequentially allocated to the *HPS*. The case of overlapped partials requires special consideration. The treatment for this case 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*. 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 *HPS*s, the overlapped partials is exclusively assigned to the one having the most dominant interpolated amplitude among all and then labeled as "usable" which means that it could be used for interpolation for its neighboring partials. For the rest of the *HPS*s the overlapped partial is labeled as existing but without a specified partial amplitude.

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

The score criteria explained in the following are designed to gracefully deal with this kind of incomplete *HPS*s. An example of treating the overlapped partials in *HPS*s of three notes is shown in Figure 3. The above plot shows the *HPS*s before the treatment and the bottom plot shows those after the treatment.

### 3.2. The score function

Having constructed the most reasonable peak sequences 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 duration *DUR* of the partial amplitude sequence, and the standard deviation of mean time *DEV*.

**Criterion 1** *HAR* is an indication of harmonicity and totally "explained" energy. It is formulated as

$$HAR = \sum_{i=1}^{I} \frac{Corr(i) \cdot Spec(i) \cdot d_M(i)}{\sum_i [Corr(i) \cdot Spec(i)]} \qquad (3)$$

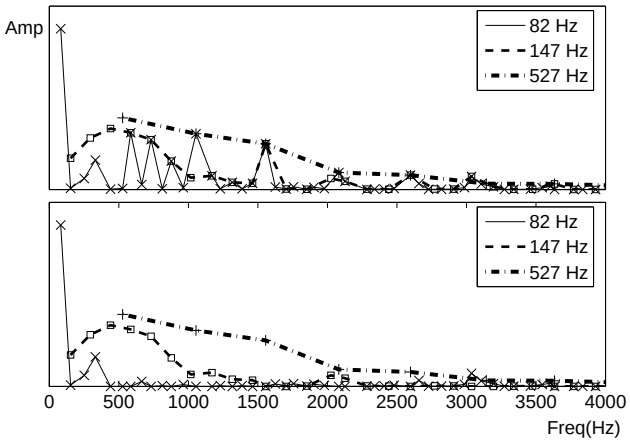


Figure 3: *Overlapped partial treatment*

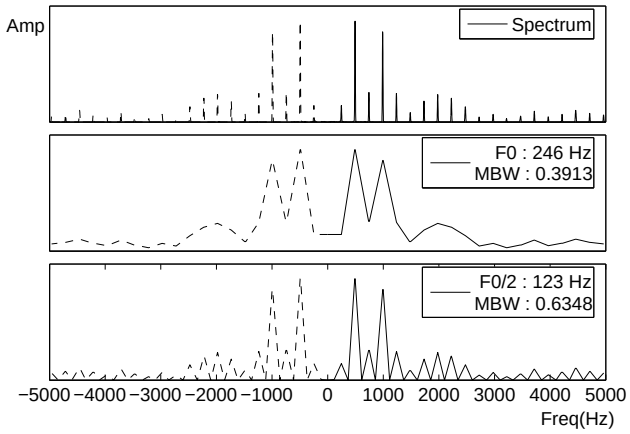


Figure 4: *Spectral smoothness comparison between F0 and F0/2*

where $I$ is the number of peaks, $i$ is the peak index, $Corr$ is the complex correlation weighting vector, $Spec$ is the linear peak amplitude and $d_M(i)$ is obtained by combining $\{d_{F0_m}(i)\}_{m=1}^{M}$ at the $i$th peak in the following way:

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

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 evaluation could perform its optimal match.

**Criterion 2** To evaluate the smoothness of a *HPS*, we calculate the mean bandwidth of the partial amplitude sequence. Each *HPS* is assembled with its "mirror sequence" to construct a new sequence $S_{F0_m}$ for further evaluation. It could also be interpreted as a hypothetical partial sequence constructed from a complex spectrum. An example of $S_{F0_m}$ is shown in the middle plot of Figure 4.

Applying $K$-point Fast Fourier Transform on $S_{F0_m}$ to obtain the linear spectral amplitude vector $X_{F0_m}$, we can calculate the mean bandwidth $MBW_{F0_m}$ as

$$MBW_{F0_m} = \sqrt{2 \cdot \frac{\sum_{k=1}^{K/2} k[X_{F0_m}(k)]^2}{\sum_{k=1}^{K/2}[X_{F0_m}(k)]^2}} \tag{5}$$

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

The function of $MBW_{F0_m}$ is to discriminate correct *F0*s from subharmonics. As the example shown in Figure 4 the spectral envelopes of a harpsichord note. Although the nature of the harpsichord does not form a smooth spectral envelope due to resonance , the *HPS* of its subharmonic *F0/2* contains even more variations and thus larger $MBW_{F0_m}$.

**Criterion 3** For a quasiharmonic sound, the spectral centroid usually lies around lower partials. Applying this general principle related to **Principle 2**, we could similarly evaluate the energy spread of the partial sequence, that is, the duration $DUR_{F0_m}$ of $HPS_{F0_m}$. Instead of removing the non-reliable components from $HPS_{F0_m}$, we simply set them to zero to maintain correct positioning of all partials. Then the duration of $HPS_{F0_m}$ could be calculated as

$$DUR_{F0_m} = \sqrt{2 \cdot \frac{\sum_{n=1}^{N_m} n[HPS_{F0_m}(n)]^2}{L \cdot \sum_{n=1}^{N_m}[HPS_{F0_m}(n)]^2}} \tag{6}$$

where $N_m$ is the length of $HPS_{F0_m}$. $L$ 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 spectral envelopes of natural sounds 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.

**Criterion 4** 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[13] and the mean time of a spectral peak can be used to characterize the amplitude evolution of the related signal[14]. For a coherent *HPS* we expect synchronous evolution resulting in a small variance of the mean time for the *HPS* of a single source.

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 variance of mean time of the partials in $HPS_{F0_m}$ is then

$$VAR_{F0_m} = \sum_{i=1}^{I}\{[\bar{t}_i - T_{F0_m}]^2 \cdot w_{F0_m}(i)\} \tag{7}$$

where $\bar{t}_i$ denotes the mean time of the $i$-th observed peak and the weighting vector $\{w_{F0_m}(i)\}_{i=1}^{I}$ is constructed by the following steps:

1) Initially set $\{w_{F0_m}(i)\}_{i=1}^{I}$ as the linear peak amplitude vector.
2) For the peaks situating too close in the observed spectrum, their spectral phases are probably disturbed. Therefore, we set the corresponding component in $\{w_{F0_m}(i)\}_{i=1}^{I}$ to 0.

3) According to the treatment of overlapped partials among $\{HPS_{F0_m}\}_{m=1}^{M}$, the components of $\{w_{F0_m}(i)\}_{i=1}^{I}$ corresponding to unusable partials are set to 0.

4) $\{w_{F0_m}(i)\}_{i=1}^{I}$ is then compressed by an exponential factor to reduce the dynamic range such that the significance of noisy peaks is raised. This makes use of noisy peaks to penalize a hypothetical partial sequence containing more noisy peaks. Finally, $\{w_{F0_m}(i)\}_{i=1}^{I}$ is normalized to be a weighting vector.

$DEV_{F0_m}$ is then defined as the square root of $VAR_{F0_m}$ divided by half of the window size.

For each combination under investigation, *MBW* of a set of *F0* hypotheses is defined as the weighted sum of $\{MBW_{F0_m}\}_{m=1}^{M}$:

$$MBW = \frac{\sum_{m=1}^{M}[\sum_{n=1}^{N_m} HPS_{F0_m}(n)] \cdot MBW_{F0_m}}{\sum_{m=1}^{M}\sum_{n=1}^{N_m} HPS_{F0_m}(n)} \quad (8)$$

This makes use of the credible components in each $HPS_{F0_m}$ as a weighting of relative importance. *DUR* and *DEV* are thus equivalently defined.

**Score function** We define the score function as

$$D_{C_M^P} = \frac{1}{\sum_{j=1}^{4} p_j}\{p_1 \cdot HAR + p_2 \cdot MBW + p_3 \cdot DUR + p_4 \cdot DEV\} \quad (9)$$

where the weighting coefficients $\{p_j\}_{j=1}^{4}$ are to be trained by an evolutionary algorithm [15]. The score function is designed in a way that smaller values stands for higher scores. Notice that *HAR* generally favors lower hypothetical *F0*s while *MBW*, *DUR* and *DEV* favor higher ones. Therefore, the criteria perform in a complementary way and the weighting coefficients should be optimized to balance the relative contribution of each criterion such that the score function generally supports correct *F0*s the best.

## 4. EXPERIMENTAL RESULTS

To evaluate the proposed *F0* estimation method, we perform a frame-based test using mixtures of musical samples. Since the criteria are designed for stationary quasiharmonic sounds, stationary parts of musical samples are pre-selected and then mixed with equal mean-square energy. 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*s.

### 4.1. Parameter optimization

The parameters to be optimized are the weighting coefficients $\{p_j\}_{j=1}^{4}$ in the score function and $\alpha$ 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[2]. Then the parameters are optimized using evolutionary algorithm and the set of parameters performing the best is used for the final evaluation on a large database.

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

### 4.2. Evaluation setups and results

Specifications for this evaluation are described below:

- Three databases: two-voice, three-voice and four-voice mixtures, labeled as TWO, THREE and FOUR respectively, are generated using McGill University Master Samples[3]. In combining $M$-voice polyphonic samples, $M$ out of twelve (C, Db, D, Eb, E, F, Gb, G, Ab, A, Bb, B) tones are preliminarily assigned and then samples ranging from 65Hz(C2) to 1980Hz(B6) are randomly selected to mix. Around 1500 samples for each database are generated in a way that each combination of note names are of equal proportion. Musical instruments not fitting the quasiharmonic model are excluded. This database contains about 30 different musical instruments. To facilitate comparison, the database is published on the first author's web page [4].
- The search range for *F0* is set from 50Hz to 2000Hz and the maximal analyzing frequency limit is fixed at 5000Hz. A Blackman window is used for analysis and all parameters are fixed for this evaluation.
- Multiple *F0* reference tables are built from single *F0* estimation of monophonic samples before mixing. A correct estimate should not deviate from the corresponding reference value by 3%. The error rates are computed by the number of error estimates divided by the total number of target *F0*s.

Evaluation using two analysis window sizes, 186ms and 93ms, are performed and the results are shown in Table 1 and Table 2, respectively. Since musical samples mixed randomly surely contain notes with harmonically related *F0*s, we present the error rates for two groups of samples: one group of mixtures containing harmonically related notes, labeled as "harmonical", and the other group "non-harmonical". The overall error rates are shown in the "total" column. The percentages of samples in the group "harmonical" are 22.43%, 32.78% and 49.46% for the three databases TWO, THREE and FOUR.

| polyphony | non-harmonical | harmonical | total |
|---|---|---|---|
| TWO | 0.58% | 7.28 % | 2.09% |
| THREE | 1.48% | 5.16 % | 2.68% |
| FOUR | 2.46% | 6.57 % | 4.50% |

Table 1: *F0* estimation results using a 186 ms window

| polyphony | non-harmonical | harmonical | total |
|---|---|---|---|
| TWO | 1.61% | 7.59% | 2.96% |
| THREE | 3.27% | 7.61% | 4.69% |
| FOUR | 5.68% | 11.78% | 8.70% |

Table 2: *F0* estimation results using a 93 ms window

The errors in the group non-harmonical are quite small which proves the satisfying performance of the proposed method. The overall errors are slightly better than the ones reported by Klapuri

[3] `http://www.music.mcgill.ca/resources/mums/html/`
[4] `http://www.ircam.fr/anasyn/cyeh/database.html`

[16], however, this comparison is not conclusive due to the fact that the testing set comprises different samples and that in [16] a larger set of samples from four different databases has been used.

## 5. DISCUSSIONS

The score function sometimes fails to correctly resolve the ambiguity concerning target *F0*s and their subharmonics or superharmonics especially *F0/2* and *2F0*. This failure scenario accounts for a great proportion of the estimation errors. Polyphonic samples mixed with musical instrument samples of rich resonances often result in this kind of wrong estimate. Taking the string instruments for example, several predominant resonances occur with the excitation [17]. If strong resonances exist in the frequency range below the fundamental, the correct *F0*s might lose too much score to subharmonics by the amount of explained energy (*HAR*). If strong resonances boost certain partials too much, correct *F0*s might lose too much score to superharmonics by the spectral smoothness (*MBW*). Dealing with resonance peaks is a key to improving robustness.

The window size is still a concern. For those mixtures containing harmonically related *F0*s, inharmonic partial structures might give a chance for correct estimation if a sufficient spectral resolution is provided. With the increase of polyphony, the performance suffers from the reduction of the window size. Therefore, investigating the techniques for treating overlapped partials is necessary.

The way of constructing polyphonic databases for evaluation should be carefully examined. With the increase of polyphony, the number of possible combinations among different notes and different instruments increases dramatically. A limited number of samples mixed in a random manner could not ensure a general representation of the large sample space. Besides, the number of harmonically related notes increases in higher polyphonic random mixtures and thus effective approaches to estimate *F0*s of exact multiple relations become more important.

## 6. CONCLUSIONS

We have presented a new method for joint evaluating the plausibility of multiple *F0* hypotheses based on three physical principles. The three principles could be interpreted as reasonable prior distribution 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. However, there are still issues to be addressed. We envisage that further improvements on the inadequate treatment for overlapped partials will lead to higher robustness.

## 7. REFERENCES


[1] 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.

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

[3] Alain de Cheveigné, "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," *Journal of Acoustical Society of America*, vol. 93, no. 6, pp. 3271–3290, 1993.

[4] Alain de Cheveigné and Hideki Kawahara, "Multiple pitch estimation and pitch perception model," *Speech Comminication 27*, pp. 175–185, 1999.

[5] Kunio Kashino and Hidehiko Tanaka, "A Sound Source Separation System with the Ability of Automatic Tone Modeling," in *Proc. of International Computer Music Conference (ICMC)*, Tokyo, Japan, 1993, pp. 248–255.

[6] Masataka Goto, "A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation using EM Algorithm for Adaptive Tone Models," in *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001)*, Salt Lake City, Utah, 2001, pp. V–3365–3368.

[7] Anssi Klapuri, "Multipitch estimation and sound separation by the spectral smoothness principle," in *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001)*, Salt Lake City, Utah, 2001.

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

[9] Axel Röbel, "Estimating partial frequency and frequency slope using reassignment operators," in *Proc. of the International Computer Music Conference (ICMC'02)*, Göteborg, 2002, pp. 122–125.

[10] Wolfgang Hess, *Pitch Determination of Speech Signals*, Springer-Verlag, Berlin Heidelberg, 1983.

[11] Thomas W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *Journal of Acoustical Society of America*, vol. 60, no. 4, pp. 911–918, 1976.

[12] H. Duifhuis and L.F. Willems, "Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception," *Journal of Acoustical Society of America*, vol. 71, no. 6, pp. 1568–1580, 1982.

[13] Loen Cohen, *Time-frequency analysis*, Prentice Hall, 1995.

[14] 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.

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

[16] Anssi Klapuri, *Signal processing methods for the automatic transcription of music, Ph.D dissertation*, Tampere University of Technology, 2004.

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