SINUSOID MODELING IN A HARMONIC CONTEXT

Wen Xue, Mark Sandler

Centre for Digital Music, Dept. Elec. Eng.,
Queen Mary, University of London, UK
{xue.wen,mark.sandler}@elec.qmul.ac.uk

ABSTRACT

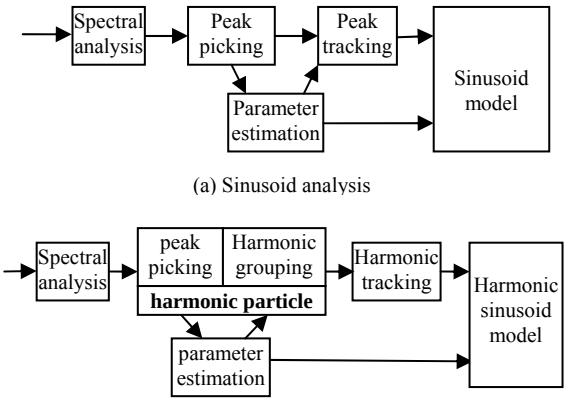
This article discusses harmonic sinusoid modeling. Unlike standard sinusoid analyzers, the harmonic sinusoid analyzer keeps close watch on partial harmony from an early stage of modeling, therefore guarantees the harmonic relationship among the sinusoids. The key element in harmonic sinusoid modeling is the harmonic sinusoid particle, which can be found by grouping short-time sinusoids. Instead of tracking short-time sinusoids, the harmonic tracker operates on harmonic particles directly. To express harmonic partial frequencies in a compact and robust form, we have developed an inequality-based representation with adjustable tolerance on frequency errors and inharmonicity, which is used in both the grouping and tracking stages. Frequency and amplitude continuity criteria are considered for tracking purpose. Numerical simulations are performed on simple synthesized signals.

1. INTRODUCTION

The standard sinusoid model [1,2] expresses an audio signal as the combination of slow-varying sinusoids plus a noise. Although the sinusoids clearly model the partials of pitched sounds, it has not been made explicit. Due to the lack of emphasis on the relationship among partials, the standard sinusoid tracking methods cannot guarantee harmonic consistence. Accordingly, the results do not provide a solid base for extracting pitched events. On the other hand, matching pursuit based methods have been proposed to extract harmonic structure from music [3, 4]. However, these methods lack the freedom of representing time-varying frequency within a single object, and tend to represent a harmonic event with time-varying pitch as multiple events. To overcome these difficulties, we apply the harmonic constraint, which is more flexible than matching pursuits, in an early stage of sinusoid analysis, preferably before the tracking of partials. This upgrades sinusoid modeling to *harmonic sinusoid modeling*. The frameworks of sinusoid and harmonic sinusoid analyzers are compared in Figure 1. The key element of the sinusoid model, the *short-time sinusoid atom*, becomes *harmonic particle*. The two main parts of the sinusoid analyzer, i.e. the sinusoid detector and the partial tracker, are replaced by harmonic particle detector and harmonic sinusoid tracker, respectively. Compared to standard sinusoid models, the harmonic model provides a higher-level description of pitched events, which enables an extensive range of analysis and synthesis operations.

A harmonic sinusoid is described by sinusoidal parameters $\{ f_l^m, a_l^m, \varphi_l^m \mid 0 \le l < L,\ 1 \le m \le M \}$, where L is the number of frames, M is the number of partials, f_l^m (a_l^m , φ_l^m) is the instantaneous frequency (amplitude, phase angle) of the m^{th} partial at the l^{th} frame. By fixing m we get a description of the m^{th} partial; by fixing l we get a description of the harmonic particle at the l^{th} frame.

This article is arranged as follows. Section 2 discusses the grouping of sinusoid atoms into harmonic particles. Section 3 discusses the harmonic sinusoid tracker. Section 4 presents some numerical results of the algorithms, followed by a brief conclusion in section 5.



(a) Sinusoid analysis

(b) Harmonic sinusoid analysis

Figure 1:*Comparing standard and harmonic sinusoid analyzers*

2. HARMONIC GROUPING WITH INEQUALITIES

We assume that the short-time sinusoid atoms have already been detected at frame *l*. This is accomplished by spectral peak picking [1] and sinusoid parameter estimation [2,5]. The harmonic grouping module collects sinusoid atoms, whose frequencies can be regarded as harmonic, from this pre-detected set. Perfect harmony is characterized by all partial frequencies being multiples of a fundamental frequency. Let f^m be the frequency of the m^{th} partial, then perfect harmony implies $f^m = mf^1$. However, this does not provide a practical way to spot harmonic particles from the pre-detected peaks, mainly for two reasons: 1) the frequency estimates carry errors, and 2) perfect harmony is not always guaranteed.

2.1. Inharmonicity

The 2nd problem, known as inharmonicity, is best known for free-vibrating strings. [6] gives an example of explicitly expressing the partial frequencies as a function of fundamental frequency f^1 and a stiffness coefficient B:

$$f^m(f^1, B) = mf^1 \cdot [1 + B(m^2 - 1)]^{1/2} \qquad (1)$$

B is a constant for a given string. Strictly speaking (1) only approximately describes the inharmonicity due to string stiffness [7], and may still carry an error. However, it is reasonable to assume that this error is so small that it can be "absorbed" into the frequency estimation error.

2.2. Frequency estimation error

The frequency estimate of a partial can be very accurate when its pitch is stable and the partial is spared of noise and disturbance. In real-world recordings, however, the pitch may have smooth or repetitive variations fast enough to affect frequency estimates, and noise and concurrent sinusoids do disturb sinusoid analyzers. The frequency estimate error depends on the estimator type and the signal behaviour, the latter being highly unpredictable. The estimation of the error bound is out of the scope of this paper. However, we always assume that we can find an error bound Δ^m for f^m. Let the frequency estimate be $\hat{f}^m$, then

$$|\hat{f}^m - f^m| < \Delta^m \tag{2}$$

The error bound Δ^m does not have to be tight. In most cases it is reasonable to set Δ^m at 1 spectral bin for low partials, and a few more for high partials if the pitch variation is fast.

Combining (1) and (2) we get

$$\hat{f}^m - \Delta^m < mf^1 \cdot [1 + B(m^2 - 1)]^{1/2} < \hat{f}^m + \Delta^m \tag{3a}$$

Equation (3) relates the frequency estimates to the two parameters of the partial frequency model (1), i.e. f^1 and B. If the frequency estimate satisfies (3a) for some f^1 and B, we allow it to be the m^{th} partial for this f^1-B pair.

2.3. Harmonic partial frequencies

Now we address the following problem: given frequency estimates of M partials, $\hat{f}^{m_1}$, $\hat{f}^{m_2}$, …, $\hat{f}^{m_M}$, where m_1^{th}, m_2^{th}, …, m_M^{th} are the partial indices, can they be grouped as harmonic partials? The answer is straightforward: if there exists f^1 and B so that (3a) holds for all the frequency estimates, then they can be regarded as harmonic partials. In other words, let the solution set of

$$\begin{cases} \hat{f}^{m_1} - \Delta^{m_1} < m_1 f^1 \sqrt{1 + B(m_1^2 - 1)} < \hat{f}^{m_1} + \Delta^{m_1} \\ \hat{f}^{m_2} - \Delta^{m_2} < m_2 f^1 \sqrt{1 + B(m_2^2 - 1)} < \hat{f}^{m_2} + \Delta^{m_2} \\ \qquad\qquad \cdots\cdots \\ \hat{f}^{m_M} - \Delta^{m_M} < m_M f^1 \sqrt{1 + B(m_M^2 - 1)} < \hat{f}^{m_M} + \Delta^{m_M} \end{cases} \tag{3b}$$

be R, then the given frequencies can be regarded as harmonic partials if and only if R≠Φ. However, (3b) is a non-linear inequality system, which makes R hard to represent in the f^1-B plane. We linearize (3b) using the substitutions

$$F = (f^1)^2, \; G = FB = B(f^1)^2 \tag{4}$$

then (3b) becomes

$$\begin{cases} g_{m_1-} < F + k_1 G < g_{m_1+} \\ g_{m_2-} < F + k_2 G < g_{m_2+} \\ \qquad\cdots\cdots \\ g_{m_M-} < F + k_M G < g_{m_M+} \end{cases} \tag{5}$$

where $g_{m-} = \left(\dfrac{\hat{f}^m - \Delta^m}{m} \right)^2$, $g_{m+} = \left(\dfrac{\hat{f}^m + \Delta^m}{m} \right)^2$, $k = m^2 - 1$. The solution of (5) is R in the F-G plane. We impose additional constraints on the allowable ranges for f^1 (e.g. 0~0.5) and B (e,g 0~0.001). These constraints are linear in the F-G plane. R, if not empty, is always a convex polygon. We represent R using a list of its N vertices in the (F-G) plane, i.e. {N; (F_n, G_n), n=0, 1, …, N-1}. To solve for R we initialize it by presetting the f^1 and B ranges (so R is a close polygon from the beginning), and apply the constraints one after another. Each constraint chops off the part of R outside a pair of parallel lines specified by $F + k_m G = g_{m\pm}$. The more partial frequencies are used, the smaller becomes R.

R represents a range for f^1 and B so that those points, and only those points in R, can be the f^1-B pairs to associate the given frequencies with. We directly have

$$\sqrt{\min_n F_n} < f^1 < \sqrt{\max_n F_n}, \; \min_n \frac{G_n}{F_n} < B < \max_n \frac{G_n}{F_n} \tag{6}$$

That is, the span of R on the F axis determines the precision of f^1, and the angular sweep of R, with respect to (0,0), determines the precision of B. The smaller R is, the more precise are f^1 and B. The m^{th} partial frequency is located by

$$f_-^m(R) < f^m < f_+^m(R), \tag{7a}$$

where

$$f_-^m(R) = m\sqrt{\min_n(F_n + kG_n)}, \; f_+^m(R) = m\sqrt{\max_n(F_n + kG_n)}. \tag{7b}$$

(7a) provides an estimation of f^m with a better precision than Δ^m. However, most of the time we need (7a) for judging whether $\hat{f}^m$ is compatible with R. If R is derived without using the m^{th} partial, then $\hat{f}^m$ can be regarded as an additional harmonic partial, as long as

$$f_-^m(R) - \Delta^{m_1} < \hat{f}^m < f_+^m(R) + \Delta^{m_1}. \tag{7c}$$

2.4. Grouping partials by harmony

The partial grouping based on the inequality representation R is simple in principle. To find a harmonic particle, one

1) initialize R, find the first partial;
2) use the found partial to update R;
3) use R to compute the range to look for the next partial;
4) find the next partial;
5) if there are still partials to find, go back to 2.

Notice that here the "first partial" refers to the first *available* partial: it does not have to be the fundamental partial, but may be any partial whose partial index is known. Figure 2 shows how R is updated for a perfect harmonic particle with neither frequency estimation error nor spurious peaks. We choose f^1=0.1 and Δ^m=0.01, 0≤B≤0.05. R obtained by using the lowest 1, 2, 3, 4 partials are shown in (a), (b), (c), (d) respectively.

In more practical cases there are three complications. First, we do not have a range to look for the first partial; second, correct partials may not appear as a spectral peak, and therefore cannot be located; third, multiple partials may be found in step 4. We discuss them in reverse order.

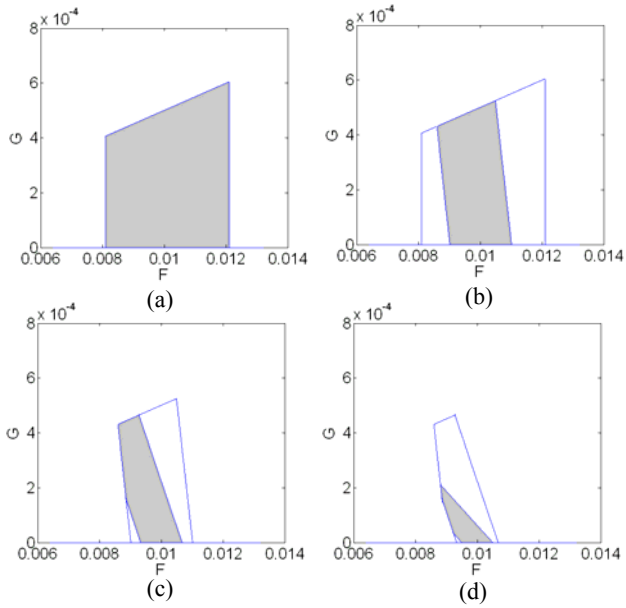


Figure 2: *Updating R using found partials*

2.4.1. Competing peaks

If more than one atom lies in the searching range for the m^{th} partial, they become competing peaks. A peak that competes with the true partial may be either a spurious peak, or a partial of a concurrent event. We can derive a *candidate* harmonic particle from every peak, and choose one from these candidate particles that is optimal in some sense. To do this we need a criterion, i.e. a scoring function, to compare two harmonic particles. The strength-harmony criterion is based on two assumptions:

(1) most spurious peaks are weak;

(2) correctly captured partials tend to have less departure from the model frequencies.

From assumption (1) we derive the strength criterion. If the strength of particle p_1 is higher than that of particle p_2, then p_1 is given a higher score on the strength side. The score can be the total amplitude calculated by summing up partial amplitudes, or the total partial SPL calculated by summing up the logarithms of partial amplitudes, or some other more perceptual measures. We always assume it can be written in an additive form, i.e.

$$s_a\left(\{\hat{a}^m\}_{m=1,2,\cdots}\right)=\sum_m s_a(\hat{a}^m),\ s_a(\hat{a}^m)\ge 0,\ s_a'(\hat{a}^m)>0. \qquad (8a)$$

Assumption (2) favours partials with more predictable frequencies. As said in 2.2, the error bounds Δ^m used for harmonic grouping are not tight. Using large error bounds provides good robustness against frequency estimation errors. However, it is a main reason that we have competing peaks. To make up for this, we introduce the harmony criterion based on the departure of frequency estimates from the model. The departure of the m^{th} partial frequency estimate $\hat{f}^m$ from model R is

$$d(\hat{f}^m,m,R)=\begin{cases} f_-^m(R)-\hat{f}^m, & \hat{f}^m<f_-^m(R)\\ 0, & f_-^m(R)\le\hat{f}^m\le f_+^m(R)\\ \hat{f}^m-f_+^m(R), & \hat{f}^m>f_+^m(R)\end{cases}. \qquad (9)$$

where $f_-^m(R)$ and $f_+^m(R)$ are defined by (7b). We also assume that the harmony score s_f can be written in an additive form:

$$s_f(\{\hat{f}^m\}_{m=1,2,\cdots},R)=\sum_m s_f^m\left(d(\hat{f}^m,m,R),s_a(\hat{a}^m)\right) \qquad (10)$$

The dependency of s_f^m on $s_a(\hat{a}^m)$ allows us to design a harmony score that is consistent with $s_a(\hat{a}^m)$ in some sense. Let Δ^m be a maximal allowable frequency departure. We choose to assign a 100% penalty to $s_a(\hat{a}^m)$ if $d(\hat{f}^m,R)\ge\Delta^m$, and no penalty if $d(\hat{f}^m,R)=0$, i.e.

$$s_f^m(d,s)=\begin{cases}0, & d=0\\ -sd/\Delta^m, & 0<d<\Delta^m\\ -s, & d\ge\Delta^m\end{cases} \qquad (11)$$

Between $d(\hat{f}^m,R)=0$ and $d(\hat{f}^m,R)\ge\Delta^m$ we assign a partial penalty, like the linear function in (11). The final score for evaluating a harmonic particle is

$$s\left(\{\hat{a}^m,\hat{f}^m,R\}_{m=1,2,\cdots}\right)=s_a\left(\{\hat{a}^m\}_{m=1,2,\cdots}\right)+s_f(\{\hat{f}^m\}_{m=1,2,\cdots},R) \qquad (12)$$

The number of candidates grows whenever we have competing peaks. However, at any stage we can combine two candidates p_1 and p_2, if 1) $s_1>s_2$ and 2) $R_1\supseteq R_2$. In particular, if the two peaks with frequency estimates $\hat{f}_1^m$ and $\hat{f}_2^m$ are competing for the m^{th} partial, $s_a(\hat{a}_1^m)+s_f^m(\hat{f}_1^m,R)>s_a(\hat{a}_2^m)+s_f^m(\hat{f}_2^m,R)$, then we can immediately discard candidate 2 if a) $\hat{f}_1^m>\hat{f}_2^m$ and $\hat{f}_1^m-\Delta^m\le f_-^m(R)$, or b) $\hat{f}_1^m<\hat{f}_2^m$ and $\hat{f}_1^m+\Delta^m\ge f_+^m(R)$. Finally the candidate harmonic particle with the highest score is selected.

2.4.2. Unfound partials

In addition to the spurious partial problem, a true partial may fail to appear as a spectral peak if 1) it is too weak, or 2) it is masked by noise. A partial being unfound is not a problem by itself, as its absence does not affect R or the searching of other partials. The real problem is that we do not know whether a partial appears as a spectral peak or not. Even when a partial does not produce a peak, it is possible for spurious peaks to appear where the partial is expected. If this were the case and the spurious peak were used to update R, the searching range of further partials would be biased. A safe way to deal with the unfound partial problem is to always reserve a candidate for "unfound partial", even when one or more atom have been located. In fact this is necessary only when the size of R is relatively large and the located atom has large frequency departure, in which case it substantially reduces the size of R. Unfound partials do not contribute to $s_a(\hat{a}_1^m)$ or $s_f^m(\hat{f}_1^m,R)$.

2.4.3. Unknown range for the first partial

The frequency range to look for a partial is calculated from R. Once the first partial has been located, R can be updated with its frequency estimates so that the search range for any further partial is reduced to no more than a few bins. This, however, does not apply to the first partial. In many cases a small frequency range of the first partial can be provided externality (s.a. by a pitch detector, a score, or a user), or by a harmonic particle in an adjacent frame during the tracking stage (see section 3). However, if there is no

such information available, we can run an exhaustive search through the pre-detected atoms. That is, we start with a strong peak, assume this is the 1st, 2nd, …, m^{th}, … partial, and derive a harmonic particle candidate from each assumption; then we compare these candidates with some criterion to choose the best one. If the audio frame has a single pitch, the found particle shall represent this pitched event. If the audio frame has multiple pitches, the found particle is interpreted as a *predominant harmonic particle*, representing one of the pitched events.

2.5. Estimating f^1 and B

The polygon R represents our knowledge of f^1 and B accumulated from the frequency estimates involved in (3b). f^1 and B do not appear explicitly during harmonic grouping or harmonic partial tracking. (6) estimates the two parameters as intervals. The sizes of the intervals are determined by the frequency estimates $\hat{f}_1^m$ and the error bounds Δ^m. As mentioned before, we use relatively large error bounds to enhance robustness. This results in a large R and imprecise f^1 and B. Accordingly, more precise estimates of f^1 and B can be obtained by reducing the overlarge error bounds. Let θ be a number between 0 and 1. By setting the error bound associated with the m^{th} partial to $\theta\Delta^m$, we can get an f^1-B range R(θ). Apparently R(1)=R, and the size of R(θ) (hence the precision of f^1 and B) is monotonous regarding θ. Since the size of R(0) is 0, we know that there exists η, 0≤η<1, so that the size of R(η) is 0, and ∀θ>η, the size of R(θ) is positive. In other words, by reducing θ from 1 to η, we shrink R(θ) from R to a zero-sized polygon. We can further prove this zero-sized polygon to be a single point. Therefore R(η) provides estimates of f^1 and B in the precise form.

We consider the constraints (3b) with argument θ. Given a point (f^1, B)∈R, it lies on R(θ) if and only if it satisfies the constraint

$$f^m(f^1,B)-\theta\Delta^m < \hat{f}^m < f^m(f^1,B)+\theta\Delta^m,\ \forall m \qquad (13a)$$

or

$$\theta > \max_m \frac{\left|\hat{f}^m - f^m(f^1,B)\right|}{\Delta^m} \equiv \theta(f^1,B) \qquad (13b)$$

θ(f^1, B) is the minimal value of θ for R(θ) to contain the point (f^1, B). We define

$$\theta(R) = \inf_{(f^1,B)\in R} \theta(f^1,B) \qquad (14)$$

θ(R) satisfies 1) for any θ<θ(R), R(θ) is empty; 2) for any θ>θ(R), R(θ) is non-empty. Therefore we have η=θ(R). The model parameters can be estimated at η:

$$(\hat{f}^1,\hat{B}) = \arg\inf_{(f^1,B)\in R} \max_m \frac{\left|\hat{f}^m - f^m(f^1,B)\right|}{\Delta^m} \qquad (15a)$$

This is a minimal-maximum problem. For the stiff string model this becomes

$$(\hat{F},\hat{G}) = \arg\inf_{(F,G)\in R} \max_m \frac{\left|\hat{f}^m - m\sqrt{F+(m^2-1)G}\right|}{\Delta^m} \qquad (15b)$$

We can then calculate f^1 and B by the inverse mapping of (4)

$$\hat{f}^1 = \sqrt{\hat{F}},\ \hat{B} = \hat{G}/\hat{F} \qquad (15c)$$

We call $e^m(F,G) = \frac{\left|\hat{f}^m - f^m(F,G)\right|}{\Delta^m}$ the relative frequency estimation error. Equations (15a) show that by shrinking R to zero-size, we locate the parameter pair that minimizes the maximal relative frequency error of all the given estimates.

The implementation of the minimal-maximum search greatly benefits from the fact that the gradient ∇e^m has constant direction. Using this property we can show that if (F,G)∈R is a local minimal-maximum of e^m, then it is also the minimal-maximum in R. In other words, the minimal-maximum of e^m is unique. A key proposition for finding the minimal-maximum is given below.

Proposition 1: if (F0, G0)∈R is not a minimal maximum, and $e_1 = e_2 = \ldots = e_K$ are K equalling maxima at (F0, G0), K>2, then there exist $l1$ and $l2$, 1≤$l1$, $l2$≤K, so that ∀1≤k≤K, along the decreasing direction of $e_{l1} = e_{l2}$, e_k - e_{l1} is non-increasing.

Proposition 1 ensures that we can always search down a curve $e_{l1} = e_{l2}$ without losing track of the maximum. The search come to a stop when there is another $l3$ so that $e_{l1} = e_{l2} = e_{l3}$. If this is not the minimal-maximum, we continue the search on curve $e_{l1} = e_{l3}$ or $e_{l2} = e_{l3}$, in the decreasing direction. It can be shown that the minimal-maximum can be reached in finite number of steps.

2.6. Detection in the presence of other harmonic particles

In polyphonic music we have multiple concurrent pitched events. It is usual that we have multiple harmonic particles in the same frame. In [7] the detection of multiple pitches is addressed as iteratively detecting and removing pitched events. In harmonic sinusoid modeling we can also detect multiple harmonic particles in a similar iterative way. Instead of removing detected events, we ignore the spectral peaks that are already collected in other harmonic particles. The harmonic grouping process remains the same. We are able to ignore certain peaks by virtue that unfound partials do not critically affect harmonic grouping. However, this makes it easier for spurious peaks to be collected. We split the harmonic grouping in two stages. In the first stage, we skip a partial whenever there is an already used peak in its searching range; in the second stage, with R already reduced to a small size, we review these skipped partials. If the used peak is still within the searching range, and it is the only peak within the range, then it is appointed to the new particle (as a *shared peak*). However, if there is another peak within the range, we take the following actions.

Let the partial index, frequency and amplitude estimates, and the f^1-B range of the used peak 1 be m_1, $\hat{f}_1$, $\hat{a}_1$ and R_1, of the unused peak 2 be m_2, $\hat{f}_2$, $\hat{a}_2$ and R_2. We define

$$s(\hat{a},\hat{f},m,R) = s_a(\hat{a}) + s_f^m(d(\hat{f},m,R), s_a(\hat{a})) \qquad (16)$$

and compare $s(\hat{a}_1,\hat{f}_1,m_1,R_1) + s(\hat{a}_2,\hat{f}_2,m_2,R_2)$ with $s(\hat{a}_1,\hat{f}_1,m_2,R_2) + s(\hat{a}_2,\hat{f}_2,m_1,R_1)$. If the former is larger, we collect peak 2 into the new harmonic particle; if the latter is larger, we replace peak 1 in the old harmonic particle with peak 2, and collect peak 1 into the new harmonic particle.

3. TRACKING HARMONIC PARTICLES

Let p_l be a harmonic particle at the l^{th} frame in time, with the f^1-B range R_l. Regarding f^1 and B of the l^{th} and $(l+1)^{th}$ frame, we assume that

$$\begin{cases} B_{l+1} = B_l \\ \left| f_{l+1}^1 - f_l^1 \right| < \Delta_l \end{cases} \tag{17}$$

(17) is the harmonic version of the frequency jump limiting used for sinusoid tracking [1]. The first line says that the inharmonicity feature remains constant during the same event, and the second line says that the pitch is not allowed to vary too fast. Given R_l and (17), we have the following inequality for the m^{th} partial at the $(l+1)^{th}$ frame:

$$f^m(f_-^1(R_l) - \Delta_l, B_l) < f_{l+1}^m < f^m(f_+^1(R_l) + \Delta_l, B_l) \tag{18}$$

This provides a range to look for any partial in the $(l+1)^{th}$ frame. To find a harmonic particle at frame $l+1$ as the successor of p_l, we initialize R_{l+1} by expanding R_l along the f^1 axis by Δ_l on both sides, i.e.

$$R_{l+1} = \{(f^1, B) \mid \exists \delta \in (-\Delta_l, \Delta_l), (f^1 + \delta, B) \in R_l\} \tag{19}$$

It can be shown that this expansion does not preserve the linearity of the sides of polygon R, so the R_{l+1} initialized strictly by (19) is no longer a polygon in the F-G plane. However, as an approximation, we can initialize R_{l+1} by expanding the vertices of R_l using (19) then take the convex hull (Figure 3). We can show that by taking this approximation R_l is expanded a little more than the amount in (19) near the sides, i.e. we are allowing a little more pitch variation at certain B's. This is not a big problem since Δ_l itself is not required to be very precise.

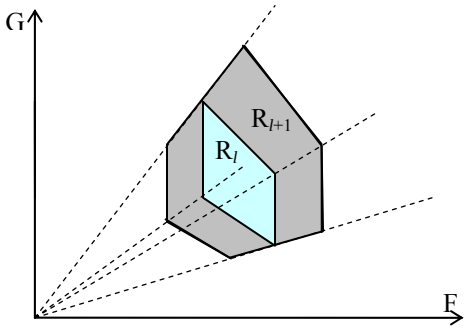


Figure 3: *Expanding R to allow pitch variation*

3.1. Short-term continuity

Once R_{l+1} has been initialized, the harmonic particle searching can be carried out using the method in section 2. However, with the knowledge of the predecessor particle, we are able to include short-term continuity criteria in the harmonic grouping stage by comparing the current candidates to the previous harmonic particle.

3.1.1. Frequency continuity

The frequency continuity has already been used to initialize R_{l+1}. However, we may have competing pitches within the allowed pitch jump. This is comparable to completing peaks in standard sinusoid modeling. In sinusoid modeling the successor is often chosen to be the peak with the smallest frequency jump [1,2] or the peak that gives the smoothest frequency contour [8, 9]. Similarly, in case of competing pitches, we choose to favour small pitch jumps. This is implemented using a pitch continuity score

$$s_p(l, l+1) = 1 - \left| \frac{f_{l+1}^1 - f_l^1}{\Delta_l} \right|^p, \; p \geq 1 \tag{20}$$

The exponent p tunes the balance of small and large pitch variations. The less is p, the less we tolerate large pitch jumps.

3.1.2. Amplitude continuity

Short-term amplitude continuity compares the partial amplitudes of candidate harmonic sinusoids of frame $l+1$ to those of frame l. We measure the similarity of two amplitude vectors $\{\hat{a}_l^m\}_{m=1,2,\cdots}$ and $\{\hat{a}_{l+1}^m\}_{m=1,2,\cdots}$ with

$$s_a(l, l+1) = \frac{2\sum_m \hat{a}_l^m \hat{a}_{l+1}^m}{\sum_m (\hat{a}_l^m)^2 + \sum_m (\hat{a}_{l+1}^m)^2} = (s_a)_1 (s_a)_2 \tag{21a}$$

$$(s_a)_1 = \frac{2\sqrt{\sum_m (\hat{a}_l^m)^2}\sqrt{\sum_m (\hat{a}_{l+1}^m)^2}}{\sum_m (\hat{a}_l^m)^2 + \sum_m (\hat{a}_{l+1}^m)^2}, (s_a)_2 = \frac{\sum_m \hat{a}_l^m \hat{a}_{l+1}^m}{\sqrt{\sum_m (\hat{a}_l^m)^2}\sqrt{\sum_m (\hat{a}_{l+1}^m)^2}} \tag{21b}$$

s_a combines two types of continuities: the total amplitude continuity $(s_a)_1$, and the amplitude distribution continuity $(s_a)_2$. $(s_a)_1$ is a measure of the change in total volume; $(s_a)_2$ is a measure of the change in short-time timbre. $0 \leq (s_a)_1 \leq 1$, $0 \leq (s_a)_1 \leq 1$.

However, we have observed that if the frequency has fast variation, the short-time amplitude continuity (21a) becomes questionable, especially when the event has a formant structure. This is due to the large variation of the short-time timbre that accompanies pitch changes. In this case we use a long-term amplitude continuity criterion, as follows.

3.2. Long-term amplitude continuity

Long-term amplitude continuity criterion is useful for events involving repetitive pitch variation. It assumes that the amplitude distributions of two frames on the same event are similar if its pitches in these two frames are close. Therefore instead of comparing the amplitude distribution with the frame closest in time, we compare it with the frame closest in frequency. Let the current harmonic sinusoid track contain frames 1, 2, …, l, with fundamental frequencies f_1^1, f_2^1, …, f_l^1, and let f_{l+1}^1 be a candidate fundamental frequency of the $(l+1)^{th}$ frame. We select $\bar{l}$ between 1 and l so that $f_{\bar{l}}^1$ is closest to f_{l+1}^1, i.e.

$$\bar{l} = \arg\min_{1 \leq k \leq l} \left| f_{l+1}^1 - f_k^1 \right|. \tag{22a}$$

We define the long-term amplitude distribution continuity score as

$$(s_a)_3 = \frac{\sum_m \hat{a}_{\bar{l}}^m \hat{a}_{l+1}^m}{\sqrt{\sum_m (\hat{a}_{\bar{l}}^m)^2}\sqrt{\sum_m (\hat{a}_{l+1}^m)^2}} \tag{22b}$$

(21a) can then be replaced by

$$s_a(l, l+1) = (s_a)_1 (s_a)_3 \tag{22c}$$

Although in (21a) and (22c) we are combining the two types of amplitude continuity measures by direct multiplication, it is not compulsory. We can use any other combination methods as long as $0 \le s_a \le 1$, with the identity $s_a=1$ for identical amplitude vectors.

3.3. Extending harmonic sinusoids

Let $p_{1,1}$, $p_{1,2}$, $p_{1,3}$, … be harmonic particles detected at frame *1*, and h_1, h_2, h_3, … be the harmonic sinusoids these particles are associated with. Now we look at the task of finding successor particles to $p_{1,1}$, $p_{1,2}$, $p_{1,3}$, …, so that h_1, h_2, h_3, … can be extended 1 frame forward. This task differs from the detection of concurrent harmonic particles (section 2.6) in that the particles detected in frame l+1 must satisfy additional continuity with the previous frame. Therefore, instead of using (12) to compare competing results, we use

$$s(l,l+1) = s_p(l,l+1) + s_a(l,l+1) \tag{23}$$

where the two addends are defined by (20) and (21a) (or (22c)). Like the detection of concurrent harmonic particles, the extension of concurrent harmonic sinusoids is performed in an iterative way, i.e. after the first harmonic particle is detected, additional harmonic particles are detected in the presence of the already found ones. The searching method remain the same, except for the scoring function (23) and the initialization of $R_{l+1,k}$ with $R_{l,k}$, k=1, 2, ….

If a successor for $p_{l,k}$ cannot be found at the $(l+1)^{th}$ frame, or any successor found for $p_{l,k}$ cannot meet a minimal continuity score, then h_k is terminated at frame l. This is the harmonic version of the *death* of a sinusoid track.

3.4. Forward harmonic sinusoid tracking

Forward harmonic sinusoid tracking creates, continues, and kills harmonic sinusoids in the forward procession of time. It takes pre-detected spectral peaks as input, and outputs harmonic sinusoids.

The forward harmonic particle tracking proceeds as follows. Let $p_{1,1}$, $p_{1,2}$, $p_{1,3}$, … be harmonic particles detected at frame 1. We associate each of them with a harmonic sinusoid, say h_1, h_2, h_3, …, i.e. $p_{1,k}$ is h_k constrained at the 1st frame. For l=2, 3, …, we do the following.

1) find the most continuous successors for the existing harmonic sinusoids (section 3.3);
   1.1) initialize $R_{l,1}$ with $R_{l-1,1}$, detect harmonic particle $p_{l,1}$;
   1.2) for k=2, 3, …, do 1.3;
      1.3) initialize $R_{l,k}$ with $R_{l-1,k}$, detect harmonic particle in the presence of $p_{l,1}$, $p_{l,2}$, …, $p_{l,k-1}$ using continuity score (23), or terminate h_k in case of failure;
2) find harmonic particles in the presence of the harmonic particles detected in 1), initialize a new harmonic sinusoid with each new harmonic particle.

3.5. Post-tracking parameter estimation

Pre-detected short-time sinusoid atoms are usually estimated using a stationary sinusoid assumption. However, accurate parameter estimation is possible only when the estimator considers parameter dynamics within a frame. Rather than estimating local dynamics from the spectrum, such as in [10], we access the dynamic information from the sinusoid tracks [11]. Post-estimation proceed in an iterative way. In each iteration, we do the following:

1) interpolate the frequency estimates using a cubic spline;
2) reestimate amplitudes using the interpolated frequency with

$$\hat{a}e^{j\hat{\varphi}} = \frac{\sum_{n=0}^{N-1} w_n^2 x_n e^{-j2\pi\int_{N/2}^{n} f(t)dt}}{\sum_{n=0}^{N-1} w_n^2} \tag{24}$$

where w_n ($0 \le n < N$) is a low-pass window function and x_n is the signal being analyzed;
3) interpolate the amplitudes using a cubic spline;
4) reestimate the frequencies by finding an approximate solution of

$$\hat{f} = \frac{\sum_{n=1}^{N-1}\sum_{m=0}^{n-1} w_{mn} a_n a_m \int_m^n f(t)dt \operatorname{sinc}\frac{\Delta\varphi_{mn}}{\pi}}{\sum_{n=1}^{N-1}\sum_{m=0}^{n-1} w_{mn} a_n a_m (n-m) \operatorname{sinc}\frac{\Delta\varphi_{mn}}{\pi}} \tag{25}$$

where $\Delta\varphi_{mn} = \int_m^n f(t)dt - 2\pi(n-m)\hat{f}$.

More details about (24) and (25) can be found in [11].

4. EXPERIMENTAL RESULTS

We run numerical tests on synthesized signals, for which the ground truth is available. The synthesized samples are 44100 points long. Amplitude and frequency variation rules include constant, exponential, and sinusoid-modulated variations. Stiff string model is applied to constant-frequency sounds. Partial amplitudes are designed to follow a 1/m rule, i.e. amplitudes are reciprocal to the partial index. We use the frame size 1024 and hop size 512. The fundamental frequency ranges from 5 bins to 40 bins (1bin=1/1024), spanning 3 octaves. We sample this range every semitone, i.e. at 37 different pitches. White noises are added to the test sampled optionally.

We measure two types of error: a harmonic grouping error and a waveform model error. The harmonic grouping error is measured by the number of correctly collected short-time sinusoid atoms divided by the total number of atoms. The waveform model error is measured by a signal-to-noise ratio, where the noise refers to the difference between the original source waveform and the resynthesized harmonic sinusoid waveform. The errors are measured independently for each test sample, which are then averaged over groups of samples.

4.1. Constant harmonic sinusoids

This group includes 925 test samples, with the 37 fundamental frequencies (f^1) from 5bins to 40 bins, 5 stiffness coefficients (B) from 0 to 0.0008, and 5 signal-to-noise ratios (SNR) from -15dB to 45dB. Given the three parameters, the test signal is synthesized by

$$M = \left\lfloor \frac{0.35}{f^1} \right\rfloor, \quad x_n = \sum_{m=1}^{M} \frac{1}{m} \cos(\varphi^m + 2\pi f^m(f^1, B)n) + r_n \tag{26}$$

The phase angles φ^m are taken at random. The noise r has been amplified to meet the selected SNR. The results are given Table 1. For stationary sinusoids the modeling is very successful, with more than 99.9% sinusoid peaks correctly collected into the partials when the SNR is above 15dB. We constantly get slightly better results for higher stiffness coefficients. This is due to the constraint of B above zero, which makes it easier to collect spurious peaks with a positive frequency departure than a negative one.

SNR / B	-15dB	0dB	15dB	30dB	45dB
0	27.5	86.4	99.9	100	100
0.0002	37.1	93.0	100	100	100
0.0004	40.6	94.2	100	100	100
0.0006	43.6	95.5	100	100	100
0.0008	44.9	95.8	99.9	100	100

(a) Group 1: peak collection rate (%)

SNR / B	-15dB	0dB	15dB	30dB	45dB
0	-0.9	14.8	30.6	45.7	60.7
0.0002	0.3	16.2	32.1	47.2	62.1
0.0004	0.6	16.5	32.4	47.5	62.3
0.0006	0.8	16.8	32.7	47.7	62.6
0.0008	1.0	17.0	32.8	47.9	62.7

(b) Group 1: Resynthesis SNR (dB)

Table 1: *Results for constant harmonic sinusoids.*

4.2. Constant pitch with exponential amplitude

Exponential amplitudes are found in real-world free vibrating bodies. This group includes 1850 test samples, with 37 fundamental frequencies (f^1) from 5 bins to 40bins, 2 stiffness coefficients (B) 0 and 0.0005, 5 amplitude decay rates (α) at -0.5, -1, -1.5, -2, -2.5 dB/frame (here "per frame" means per hop size, i.e. per 512 points) , and 5 SNRs from -15dB to 45dB. Given the four parameters, the test signal is synthesized as

$$x_n = \sum_{m=1}^{M} \frac{10^{\alpha m/10240}}{m} \cos(\varphi^m + 2\pi f^m(f^1, B)n) + r_n \qquad (27)$$

where M, φ^m and r are determined in the same way as in (26). The results are given in Table 2.

SNR / α	-15dB	0dB	15dB	30dB	45dB
-0.5	28.1	78.5	99.2	100	100
-1	21.5	55.4	84.5	98.9	100
-1.5	17.4	40.6	63.6	84.7	96.7
-2	14.8	32.7	49.6	67.7	81.6
-2.5	13.7	27.6	41.8	55.4	69.0

(a) Group 2: peak collection rate (%)
α: amplitude decay rate (dB/frame)

SNR / α	-15dB	0dB	15dB	30dB	45dB
-0.5	-0.2	15.2	31.0	46.3	61.2
-1	-0.5	13.9	30.0	45.6	59.7
-1.5	-0.9	12.9	21.7	44.5	56.1
-2	-1.3	11.8	23.7	43.2	49.3
-2.5	-1.8	12.5	21.9	26.7	22.0

(b) Group 2: Resynthesis SNR (dB)

Table 2: *Results for exponential amplitudes.*

The decay rate has a very regular effect on both errors, partially because the signal drops below noise level after certain points. Although in this test all partials have the same decay rate, for partial-dependent decay rates, which is common in real music signals, the behaviour is similar: all partials that falls below the noise level become hard to pick up. Unlike matching pursuits, sinusoid modeling does not assume any specific coupling between partial amplitudes.

4.3. Constant pitch with modulated amplitude

This group includes 550 samples, with 22 fundamental frequencies (f^1) from 5bins to 40bins (3 octaves on diatonic scale) , 5 modulation depths (d) 0.1, 0.2, …, 0.5, 5 modulator periods (T) 2, 4, …, 10 frames, SNR is fixed at 15dB. Given the four parameters, the test signal is synthesized as

$$x_n = \sum_{m=1}^{M} \frac{1}{m}\left(1 + d\cos\frac{\pi n}{256\mathrm{T}}\right)\cos(\varphi^m + 2\pi m f^1 n) + r_n \qquad (28)$$

where M, φ^m and r are determined in the same way as in (26). The results are given in Table 3.

d \T	all
all	> 99.98%

(a) Group 3: peak collection rate

d \ T	2	4	6	8	10
0.1	28.17	30.34	30.55	30.57	30.60
0.2	24.64	29.57	30.36	30.56	30.56
0.3	21.85	28.60	30.15	30.42	30.49
0.4	19.74	27.54	29.77	30.31	30.44
0.5	18.09	26.58	29.48	30.17	30.39

(b) Group 3: Resynthesis SNR (dB)
d: modulation depth; T: modulator period (frames)

Table 3: *Results for exponential amplitudes.*

With the SNR at 15dB, the partial collection rate stays consistently close to 100%. The waveform error increases with modulation depth and frequency.

4.4. Pitch modulation with constant amplitudes

This group includes 550 samples, with 22 fundamental frequencies (f^1) from 5bins to 40bins (3 octaves on diatonic scale), 5 modulator amplitudes (d) 0.3, 0.6, …, 1.5 semitones, 5 modulator periods (T) 2, 4, …, 10 frames, SNR ratio is set to 15dB. Given the four parameters, the test signal is synthesized as

$$x_n = \sum_{m=1}^{M} \frac{1}{m}\cos\left(\varphi^m + 2\pi m \sum_{l=0}^{n-1} f^1\left(1 + \left(2^{\frac{d}{12}} - 1\right)\cos\frac{\pi l}{256\mathrm{T}}\right)\right) + r_n \qquad (29)$$

where M, φ^m and r are determined in the same way as in (26). Again the partial collection rate stays consistently close to 100%. We list the resynthesis SNR's in Table 4. Only amplitude reestimation is used in the post-tracking stage to generate these results.

d \ T	2	4	6	8	10
0.3	14.5	23.6	27.9	29.0	29.3
0.6	10.8	17.9	21.5	25.4	27.0
0.9	7.7	14.7	17.7	21.3	24.0
1.2	6.0	11.2	13.0	18.5	21.0
1.5	4.8	8.3	7.8	13.0	18.9

Group4: Resynthesis SNR (dB)
d: modulator amplitude (semitones); T: modulator period (frames)

Table 4: *Results for vibrato.*

If we compare Table 4 with Table 3(b), we see that a frequency modulation of as small as 0.3 semitones brings more error than an amplitude modulation of 50% the central value.

5. CONCLUSIONS

In this article we have proposed a harmonic sinusoid modeling system, and discussed the harmonic sinusoid analyzer in brief. The harmonic sinusoid model is an update to the standard sinusoid model. Unlike sinusoid models that describe mostly low-level spectral contents, harmonic sinusoids directly model pitched events, which could provide solid starting points for music-related tasks. An application of this model in audio editors has been proposed in [12].

The current model can be further improved on several aspects. 1) The partial harmony has its origin in 1-dimension simple harmonic oscillation in string and air column, and does not describe membrane or bar vibration, which lies behind percussion instruments such as the kettledrum and marimba [13]. The analysis of these instruments requires partial frequency coupling rules different from simple harmony. 2) Even for harmonic instruments, there may exist extra partials that do not fall within a harmonic context [14]. These can be picked up by introducing individual spectral lines into the model, or be included in a more comprehensive harmonic model. 3) Harmonic tracking can be further refined by introducing finer frequency and amplitude continuity criteria, and the use of object models in partial tracking. 4) In [12] we have proposed the use of forward-backward searching [15] where atoms can be located at multiple frames, so that the tracking is more robust to local disturbance. 5) The current model treats very close (or overlapping) partials from two or more harmonic sinusoids as a shared partial; we also need separation techniques to resolve these shared partials into individual harmonic sinusoids. 6) On the synthesizer side, a more robust and accurate modeling of time-varying sinusoids is necessary to achieve better SNRs.

6. ACKNOWLEDGEMENTS

This work was supported by EPSRC EP/E017614/1 project OMRAS2 (Online Music Recognition and Searching) and Centre for Digital Music.

7. REFERENCES

[1] R. J. McAulay, T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation，" *IEEE Tran. ASSP*, vol. 34, pp. 744-754, 1986.

[2] X. Serra, Musical Signal Processing, chapter Musical Sound Modeling with Sinusoids plus Noise, pp. 91–122, G. D. Poli and A. Picialli and S. T. Pope and C. Roads Eds. Swets & Zeitlinger, 1996.

[3] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE Tran. Signal Proc.*, vol.51 no.1, 2003.

[4] C. Duxbury, N. Chetry , M. Sandler and M. E. Davies , "An efficient two-stage implementation of harmonic matching pursuit," in *Proc. EUSIPCO04*, Vienna, 2004.

[5] F. Keiler, S. Marchand, "Survey on extraction of sinusoids in stationary sounds," in *Proc. DAFx'02*, 2002, pp.51-58.

[6] A. Klapuri, "Wide-band pitch estimation for natural sound sources with inharmonicities," in *Proc. AES 106th Convention*, Munchen, 1999.

[7] A. Klapuri, T. Virtanen, J.-M. Holm, "Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals," in *Proc. DAFx'00*, 2000.

[8] M. Lagrange, S. Marchand, M. Raspaud, J.-B. Rault, "Enhanced partial tracking using linear prediction," in *Proc. DAFx'03*, 2003.

[9] P. Depalle, G. Garcia, X. Rodet, "Tracking of partials for additive sound synthesis using hidden Markov models," in *Proc. ICASSP*, Minneapolis, 1993.

[10] A. Röbel, "Estimating partial frequency and frequency slope using reassignment operators," in *Proc. ICMC*'02, Göteborg, 2002.

[11] Wen X., M. Sandler, "Error compensation in modeling time-varying sinusoids," in *Proc. DAFx'06*, Montreal, 2006.

[12] Wen X., M. Sandler, "New audio editor functionality using harmonic sinusoids," in *Proc. AES 122nd Convention*, Vienna, 2007.

[13] N. H. Fletcher and T. D. Rossing, The Physics of Musical Instruments (2nd Edition), Springer-Verlag New York, Inc., 1998.

[14] H. A. Conklin, "Generation of partials due to nonlinear mixing in a stringed instrument", *J. Acoust. Soc. Am.*, vol.105, no.1, January 1999, pp.536-545.

[15] S. Austin, R. Schwartz, P. Placeway, "The forward-backward search algorithm," in *Proc. ICASSP91*, Toronto, 1991.