NUMBER THEORETICAL MEANS OF RESOLVING A MIXTURE OF SEVERAL HARMONIC SOUNDS

Anssi Klapuri
Signal Processing Laboratory, Tampere University of Technology,
P.O.Box 553, FIN-33101 Tampere, FINLAND
Tel: +358 3 3652124; fax: +358 3 3653857
e-mail: klap@cs.tut.fi

ABSTRACT

In this paper, a number theoretical method is developed for the purpose of analyzing the spectre of a mixture of harmonic sounds. The method is based on the properties of prime numbers and on non-linear filtering. It is shown that a number theoretical approach is of vital importance in order to detect and observe harmonic sounds in musical polyphonies. The method is verified by applying it to the automatic transcription of piano music.

1 INTRODUCTION

Multiple fundamental frequency tracking is an almost unexplored area of research, although in the moniphonic case several algorithms have been proposed that are robust, commercially applicable and operate in real time. Some published efforts towards multipitch tracking have been made in the field of automatic transcription of music [1,2]. Until these days, however, the performance of the transcription systems has been very limited in polyphonic signals.

We will discuss the spectral properties of a mixture of harmonic sounds and demonstrate why single pitch tracking algorithms are not appropriate as such for use in polyphonic signals. Then we attempt to establish a number theoretical method to detect and observe harmonic sounds in polyphonic signals. This does not only concern multiple fundamental frequency tracking, but observing any of the features of harmonic sounds in polyphonic signals.

2 *FEATURE* OF A SOUND

A harmonic sound consists of a series of frequency partials, harmonics. They appear as peaks in the frequency spectrum at constant frequency intervals $f0$, with the lowest partial at frequency $f0$, which is therefore called the fundamental frequency of the sound.

We denote harmonic sounds with uppercase letters *S* and *R*. These are used consistently in such roles that sound *S* is being observed in the interference (presence) of a sound *R*, or R_i, if there are several interfering sounds. We denote the harmonic partials of a sound by h_j, where $j \geq 1$. Braces are used to denote *sets*, thus $\{h_j\}$ being a set of harmonics.

Further, we denote by $g(x)$ a *feature* of *x*, where *x* can be a sound *S*, or its single harmonic partial h_j. We will separate the different features by subscript characters, for example $g_F(x)$, $g_L(x)$ and $g_T(x)$ refeffing to the frequency, loudness, and onset time of *x*, respectively. Because the very substance of a harmonic sound is its series of equidistant sinusoid partials, any observation of a harmonic sound must rely on its harmonic partials, no matter if it is made in time or in frequency domain.

3 BASIC PROBLEM IN RESOLVING A MIXTURE OF HARMONIC SOUNDS

There are several good methods for measuring the frequency and amplitude contours and phases of the sinusoid partials in a signal [3,4,5]. Separating a *mixture* of harmonic sounds is problematic for two specific reasons.

1. It is most difficult to organize sinusoid partials to their due fundamental frequencies, because most often the harmonic series of different sounds extend to common frequency bands.
2. The amplitude envelopes and phases of two sinusoids can no more be deduced from their sum, if they *overlap*, i.e. share the same frequency.

Proposition 1. If any harmonic h_j^S of a sound *S* is overlapped by any harmonic h_i^R of an interfering sound *R*, then the fundamental frequency of the sound *R* must be $f0_R = \frac{m}{n} \cdot f0_S$, where *m* and *n* are positive integer numbers.

Proof. The condition of a harmonic h_j^S of a sound *S* to be overlapped by a harmonic h_i^R of an interfering sound *R* can be expressed as

$$i \cdot f0_R = j \cdot f0_S . \qquad (1)$$

When the common factors of *j* and *i* are reduced, this can be expressed as

$$f0_R = \frac{m}{n} \cdot f0_S , \qquad (2)$$

where $(m, n) \geq 1$ and can be calculated from the integers *i* and *j*. ❑

Proposition 2. If the fundamental frequencies of two harmonic sounds *S* and *R* are $f0_S$ and $f0_R = \frac{m}{n} \cdot f0_S$, respectively, then every n^{th} harmonic h_{nk} of the sound *R* overlaps every m^{th} harmonic h_{mk} of the sound *S*, where integer $k \geq 1$.

Proof. Substituting (2) to (1) we can rewrite the condition of a harmonic h_j of a sound *S* to be overlapped by a harmonic h_i of an interfering sound *R* as

$$(i \cdot f0_R = j \cdot f0_S) \Leftrightarrow \left(i \cdot \frac{m}{n} \cdot f0_S = j \cdot f0_S \right) \Leftrightarrow (i \cdot m = j \cdot n) ,$$

which is true for each pair i=nk and j=mk, where $k \geq 1$. ❑

It is easy to see that if *m*=1 in equation 2, *R* overlaps all the harmonics of *S* at their common frequency bands. In this case, detecting and observing *S* is difficult and even theoretically ambiguous. This case will be separately discussed.

4 CERTAIN PRINCIPLES IN WESTERN MUSIC

An important principle governing music is paying attention to the frequency relations, intervals, of simultaneously played notes. Two notes are in a harmonic relation to each other if their fundamental frequencies satisfy

$$f0_2 = \frac{m}{n} \cdot f0_1, \quad (3)$$

where m and n are small integers. The smaller the values of m and n are, the closer is the harmonic relation of the two sounds and the more perfectly they play together.

Western music arranges notes to a quantized logarithmic scale, where the fundamental frequency of a note k is $f0_k = 440 \cdot 2^{k/12}$ Hz, and $-48 \le k \le 39$ in a standard piano keyboard, for example. Although the scale is logarithmic, it can surprisingly well produce the different harmonic intervals that can be derived from Equation (3) by substituting small integers to m and n. The realizable musical intervals deviate a little from their ideals, but the amount of error is so little that it practically does not disturb the human ear. Moreover, for a feasible frequency analysis resolution, the overlapping of the harmonics of the two sounds is the same as if the harmonic relation were perfect.

For instance, the fundamental frequencies of the notes in a basic *major* chord are in $4 : 5 : 6$ relations to each other. Based on the proposition 2, 47%, 33% and 60% of the harmonic partials of the notes are overlapped by the other two notes in the chord. In this case, 60% of the partials of the third note would be found from the signal even in its absence. This demonstrates why the algorithms that have been designed for the detection and observation of a single harmonic sound cannot be straightforwardly applied to resolving polyhonic musical contents. Instead, we need to rethink the very kernel, how to collect the information of a sound from its harmonics.

5 PRIME NUMBER HARMONICS

Prime number harmonics $\{h_1, h_2, h_3, h_5, h_7, ...\}$ of a sound share a desired common property that is derived from the very definition of the prime numbers: they are divisible only by one and themselves. This has an important consequence, which will give a steadfast starting point in organizing frequency partials to their due fundamental frequencies.

Proposition 3. Any harmonic sound R can overlap only one prime number harmonic of a sound S, provided that the fundamental frequency of R is not $f0_R = \frac{1}{n} \cdot f0_S$, where integer $n \ge 1$. If R overlaps two prime number harmonics of S, it overlaps all the harmonics of S, and its fundamental frequency is in the mentioned relation to S.

Proof. This can be proved by assuming that two prime number harmonics of S are overlapped by the harmonics of R and showing that in this case $f0_R = \frac{1}{n} \cdot f_S$, where $n \ge 1$, and the sound R overlaps all the harmonics of the sound S.

Let $f0_S$ and $f0_R$ be the fundamental frequencies of the sounds S and R, respectively. We denote an arbitrary prime number by p_i. The condition of two prime number harmonics of S being overlapped by any harmonics h_j of R can be expressed as

$$\begin{cases} i_1 \cdot f0_R = p_1 \cdot f0_S \\ i_2 \cdot f0_R = p_2 \cdot f0_S \end{cases}, \quad (4)$$

where p_2 can be solved as

$$p_2 = \frac{p_1 \cdot i_2}{i_1}.$$

In order for p_2 to be a prime number and not equal to p_1, i_1 must satisfy

$$i_1 = n \cdot p_1, \quad (5)$$

where n is an integer and implies

$$i_2 = n \cdot p_2.$$

Substituting (5) to (4) we get

$$f0_R = \frac{p_1 \cdot f0_S}{i_1} = \frac{p_1 \cdot f0_S}{n \cdot p_1} = \frac{f0_S}{n}, \quad (6)$$

where $n \ge 1$. ❑

If Equation 6 holds, all the harmonics of S are overlapped by every n^{th} harmonic of R, based on proposition 2.

6 DEALING WITH *OUTLIER* VALUES

Let us denote the set of prime harmonics by $\{h_p \mid p \text{ is prime}\}$, and the set of the features of the prime harmonics by $\{g(h_p) \mid p \text{ is prime}\}$, where the type of the feature is not yet fixed. Based on proposition 3, prime number harmonics of a sound S can be considered as independent pieces of evidence for the existence of the sound S, or for any of its features that can be deduced from its harmonics.

In the set of representative features $\{g(h_p) \mid p \text{ is prime}\}$ there are two kinds of *outliers*, i.e., irrelevant values in respect of the true feature $g(S)$ of the sound. Some prime harmonics have been disturbed by interfering sounds, while others may be totally lacking from S. Those values that are not outliers vary somewhat in value, but outliers are single, clearly deviated values, and invalid to represent the true feature of S. However, a majority of the representatives should be reliable, it being unprobable that a majority of the prime number harmonics would be either missing or each corrupted by an independent interfering sound.

This is the motivation for the design of a filter which would pick the estimated feature $\hat{g}(S)$ from the set of independent representatives $\{g(h_p) \mid p \text{ is prime}\}$ and drop out the irrelevant values. The class of median and order statistic filters is prompted by the fact that they are particularly effective in dealing with the kind of data that was characterized above. These filters depend on *sorting* the set of representatives. Under or overestimated outlier values map to the both ends of the sorted set, and in between, the reliable samples are sorted from the smallest up to the largest value. Thus a trivial way to estimate a feature of a sound would be

$$\hat{g}(S) = median\{g(h_p) \mid p \text{ is prime}\}. \quad (7)$$

Weighted order statistic (WOS) filters are defined in [7]. They allow convenient tailoring of the filter's sample selection probabilities. The j^{th} *sample selection probability* is the probability that the sample h_j in a set $\{h_j\}$ is selected to be the output of the filter [8]. We denote the sample selection

probabilities of a filter by $P_s(j)$.

7 GENERALIZATION OF THE RESULT

A still remaining shortcoming of the proposed procedure is that it utilizes only the prime number harmonics. This degrades the usability of the algorithm and makes it sensitive to the tonal content of a sound. We proceed towards a model where this defect is removed but the advantages of the set of prime number harmonics are preserved.

We denote by v a WOS filter that picks the estimated feature of a sound from the set of features of its harmonics. This can be written as

$$\hat{g}(S) = v\{g(h_j)\} . \quad (8)$$

Further, we denote by $E_m = \{h_{mj}\}$, $j \geq 1$, a set which contains every m^{th} harmonic of a sound, starting from harmonic m. In Proposition 2 we proved that if an interfering sound R overlaps a harmonic of an observing sound, it overlaps every m^{th} harmonic of it, i.e., exactly the subset E_m.

The requirements of the filter v can now be exactly expressed as follows. Given a number N of interfering sounds, they should together contribute only up to a limited probability λ that a corrupted harmonic is chosen to the output of v. At the same time, the filter should utilize all the harmonics of the observed sound as equally as possible to make it applicable and robust to different kinds of sounds.

These requirements can be achieved by finding sample selection probabilities $P_s(j)$ for the filter v so that the selection probabilities of the N largest subsets E_m together sum up to the given limit probability λ. N largest sets that are not subsets of each other are the prime sets $\{E_m \mid m=2,3,5,7...\}$. E_1 is excluded since the case of all harmonics being overlapped will be discussed separately. If N is set to 1 this can be expressed as finding $P_s(j)$ in a minimizing problem

$$min\left\{ \max_{m \geq 2} \left\{ \sum_{i=1}^{J} P_s(m \cdot i) \right\} \right\}, \quad (9)$$

where J denotes the total number of detectable harmonics of the observed sound.

We assume all fundamental frequencies of interfering sounds R_i to be equally probable. Based on the assumption, all m and n values binding $f0_R$ in equation 2 are equally probable, from where it follows that R is equally probable to choose to overlap any subset E_m. However, the relative trustworthiness is not the same for all the single harmonics h_j, but equals the probability that none of the sets E_m that h_j belongs to is overlapped. This is calculated as $\tau^{D(j)}$, where τ represents the overall probability of an interfering sound to overlap some subset E_m and $D(j)$ is the number of subsets E_m that harmonic h_j belongs to. It can be easily proved that $D(j)$ is the number of integers that divide j, $D(1)=1$. An integer a is defined to *divide* another integer b, if and only if $b = da$ holds for some integer d [9].

Selection probabilities $P_s(j)$ of the harmonics should be according to their probability of being trustworthy. We can therefore write $P_s(j)$ in the form

$$P_s(j) = \tau^{D(j)} , \quad (10)$$

where $j \geq 1$, and $D(j)$ is as defined above.

We can now rewrite the requirements of the feature exraction filter v as

$$\sum_{j \in I} \tau^{D(j)} = \lambda \cdot \sum_{j=1}^{J} \tau^{D(j)} , \quad (11)$$

where set I is defined to contain the numbers j of the harmonics h_j that belong to some of the N largest subsets $\{E_m \mid m=2,3,5,7...\}$. If $N=1$, set I simply contains even numbers up to J. Thus the left side sums the selection probabilities of the harmonics in the N largest subsets. The right side summation goes over the selection probabilities of all the harmonics and should equal unity.

From equation 11, τ can be solved. If the problem is solvable for given N, λ and J, there is only one root that is real and between 0 and 1. This root is the earlier discussed value of τ. Selection probabilities $P_s(j)$ can then be calculated by substituting τ to Equation 10, and scaling the overall sum of $P_s(j)$ to unity.

We arrive at selection probabilities $P_s(j)$, where N interfering sounds may together contribute only up to λ probability that an overlapped harmonic exists in the output. Another very important property of this algorithm is that we can flexibly make a tradeoff between the two requirements of the filter: the less we put emphasis on the robustness of the filter v in the presence of interfering sounds, the more equally the filter utilizes all the harmonics of the observed sound, and vice versa. Figure 1 illustrates the selection probabilities for N=2, λ=0.45 and J=20.

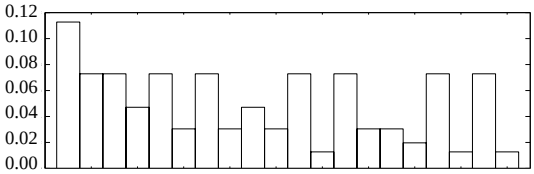

Figure 1. $P_s(j)$ for 20 harmonics, when N=2 and λ=0.45.

Thus we reduced the observation of a feature $g(S)$ of a harmonic sound S in the presence of other interfering harmonic sounds R_i to measuring the features $g(h_j)$ of the harmonics of the sound S and applying a weighted order statistic filter v to yield an estimate for $g(S)$. A design procedure to find a WOS filter which *implements* the calculated selection probabilites is presented in [10].

8 FEATURE SUBTRACTION PRINCIPLE

Our algorithm and discussion on the observation of the features of a harmonic sound in the presence of other harmonic sounds was based on an assumption that the observed sound S is not *totally* overlapped by an interfering sound R, whose fundamental frequency is $f0_R = \frac{1}{n} \cdot f0_S$.

The basic idea of our solution to this problem is to *compensate* the effect of the interfering sound R, the properties of which can be robustly extracted in the presence of S using the procedure presented before, because the interfering sound is not totally overlapped by S. Thus it will be enough to develop an algorithm to subtract, remove, or compensate, the revealed properties of the lower sound and then proceed

to determine the properties of the sound *S*, which is laid bare from under the interfering sound. The subtraction process depends on the feature under inspection, and cannot be presented in a general form.

9 ALGORITHM EVALUATION

Our algorithm was evaluated and verified by applying it in a computer program whose purpose is to transcribe polyphonic piano music. The program is first allowed to study piano notes one by one, a sufficient amount to represent all the different *tone colours* that can be produced by that instrument. After this we require the program to transcribe rich polyphonic musical signals played with the same instrument, i.e., to determine the fundamental frequencies and the loudnesses of the sounds in the signals.

In all test cases, the transcription was done without knowledge of the polyphony of the transcribed signals, and with a fixed constant set of parameters. The range of fundamental frequencies was restricted to extend from 65 Hz to 2100 Hz, where five octaves and 61 piano keys fit in between. An acoustic upright piano was used in simulations.

Transcription proceeds by first detecting all potential note candidates in the spectrum, and then resolving their loudnesses one by one, using the new method. Naturally, there are much more potential note candidates than truly played notes. We call *true notes* the notes that were truly played in the recorded signals, and *false notes* those that appear as note candidates, although they were not actually played.

The goodness of the algorithm is justified by its ability to indicate the truly existing sound in the signal, i.e. the loudness of the true notes should raise clearly above the loudness of the false ones. The loudnesses of the candidates in each time segment are scaled between the values 0 to 100.

Certain note combinations were separately played and fed to the transcriber to review its ability to resolve rich musical polyphonies. Results are presented in Table 1. In the first type of tests, consonant and dissonant chords were played, two in each of the five octaves. In the second test, groups of adjacent notes in the piano keyboard were played in each octave. Third, groups of seven random notes were allotted in the allowed range of pitces and played. In each of these tests, average loudnesses among the true and false notes were calculated and recorded. Also the worst cases, where the loudness of the false notes gets closest to the true notes, was recorded. The polyphony, number of notes in each test, is also indicated.

Table 1: Relative loudnesses of the true and false notes.

Test type	Polyp.	Averages		Worst case	
		min true	max false	min true	max false
chords	3-4	64	17	78	43
groups	4-5	48	7	30	7
random	7	21	11	10	11

Chosen classical compositions were played and excerpts from them were posed to our transcription system to test its practical transcription efficiency. Here we used 25% relative loudness as a threshold to segregate between true and false notes. Weaker candidates were discarded as false notes. Results are presented in Table 2. The last piece was played by a computer on an electric piano. The effect of all notes having roughly equal playing loudness and the absence of cross resonance and noise can be noticed in the results.

Table 2: Transcription results using 25% loudness limit.

Composition	Notes in total	Typical polyphony	Missing notes	Erroneous extra notes
Für Elise, I	86	1	-	3
Für Elise, II	190	half:2 half:4	9	6
Inventio 8	205	2	15	4
Rondo alla Turca	142	3 (up to 5)	1	-

Finding the exact fundamental frequencies of the sounds in the analyzed signals proved successful in all cases. They were not assumed to be quantized to the closest legal notes.

10 CONCLUSION

The problem of resolving rich musical polyphonies was the motivation for developing the new methods. Simulations illustrate that the current system works within certain error limits up to seven notes polyphony. Especially, although increase in polyphony brings the levels of the weakest true note and the strongest false note closer to each other, the system does not totally break down even in rich polyphonies. We conclude that a number theoretical analysis of a sound mixture is the key to a robust detection and observion of harmonic sounds in the interference of each other.

REFERENCES

[1] Kashino, Nakadai, Kinoshita, Tanaka. “Application of Bayesian probability network to music scene analysis”. Proceedings of the Int. Joint Conference on Artificial Intelligence, CASA workshop, 1995

[2] Martin. “A Blackboard System for Automatic Transcription of Simple Polyphonic Music”. MIT Media Laboratory Perceptual Computing Section Technical Report No. 399, 1996.

[3] McAulay, Quatieri. “Speech analysis/synthesis based on a sinusoidal representation”. *IEEE Trans. ASSP*, 34(4), pp. 744-754, 1986.

[4] Depalle, García, Rodet. “Tracking of Partials for Additive Sound Synthesis Using Hidden Markov Models”. *IEEE Trans. on ASSP*, 1993.

[5] Serra. “Musical Sound Modeling With Sinusoids Plus Noise”. Roads, Pope, Poli (eds.). “Musical Signal Processing”. Swets & Zeitlinger Publishers, 1997.

[6] Astola, Kuosmanen. “Fundamentals of Nonlinear Digital Filtering”. CRC Press LLC, 1997.

[7] Kuosmanen. “Statistical Analysis and Optimization of Stack Filters”. Tech.D. thesis., Acta Polytechnica Scandinavia, Electrical Engineering Series, 1994.

[8] Koblitz. “A Course in Number Theory and Cryptography”. Springer, Berlin, 1987.

[9] Klapuri. “Automatic Transcription of Music”. MSc thesis, Tampere University of Technology, 1998.

11 THRASHCANNED (Did not fit to the paper)

output of a WOS filter is calculated by repeating each input sample up to its weight, sorting the resulting multiset, and then choosing the T^{th} smallest value from the resulting multiset. The weights and the threshold T are the parameters of the filter. The weights of a WOS filter

To illustrate this we shortly review the results of an older reference system which was based on straightforward pattern recognition. Before the presented algorithm was recovered, we developed a system that was aimed to solve the same problem as the one presented earlier

To evaluate the efficiency of the transcriber in *giving prominence to the true notes*, we represent the results before the final segregation which will . Thus the loudness of the true note candidates should raise clearly above the loudness of the false ones. Loudnesses are represented in relative terms, so that the loudness of the loudest note is always 100 %. It would be desirable to keep loudnesses of the true notes above 40 % and that of false notes below 20 %.

We will also compare this transcription system to our older pattern recognition approach, reflect their differences and strengths, and evaluate the role of the new methods.

It turned out, however, that tone patterns in the straightforward pattern recognition approach are completely sunk under each other already in three notes polyphony. This is due to the properties of an overlapping harmonic series and the percentages of the overlapped amounts that were discussed in the previous chapter.bc

[10] We denote the total number of detectable harmonics of the observed sound by J. The overall selection probability of the harmonics is

$$\sum_{j=1}^{J} \tau^{D(j)} \qquad (12)$$

and should equal to unity.

Set I is defined to contain the numbers j of the harmonics h_j that belong to some of the N largest subsets $\{E_m \mid m=2,3,5,7...\}$. If N=1, set I simply contains even numbers up to J. The sum over the selection probabilities of the harmonics in these N largest subsets is

$$\sum_{j \in I} \tau^{D(j)} \,. \qquad (13)$$

Finally, we can rewrite the first requirement of the feature exraction filter *v* (see Table x on x) as

$$\sum_{j \in I} \tau^{D(j)} = \lambda \cdot \sum_{j=1}^{J} \tau^{D(j)} \,, \qquad (14)$$

from which τ can be solved. Although this cannot be done analytically for higher than fourth order polynomials, efficient numerical methods exist [Horn85]. In practice, this reduces to finding eigenvalues of the associated companion matrix of A [Horn85].

If the problem is solvable for given N, λ and J, there is only one root that is real and between 0 and 1. This root is the earlier discussed value of τ. Selection probabilities $P_s(j)$ can now be solved by substituting τ to Equation 10.