Multi-Channel Excitation/Filter Modeling of Percussive Sounds with Application to the Piano

Jean Laroche
Télécom Paris
46 Rue Barrault
75634 Paris Cedex 13
FRANCE
laroche@sig.enst.fr
Tel: (33 1) 45 81 78 62
Fax: (33 1) 45 88 79 35

Jean-Louis Meillier
Laboratoire d'Acoustique de l'Université du Maine
Route de Laval B.P. 535
72017 Le Mans Cedex
France

Abstract— **This paper discusses models for sounds of percussive musical instruments such as vibraphones, pianos, and the like. In addition to classical source/filter models a 'multi-channel excitation/filter model' is proposed in which a single excitation is used to generate several sounds, for example five piano tones belonging to the same octave. Techniques for estimating the model parameters are presented along with their application to the sound of a real piano. Our experiments demonstrate that it is possible to calculate a single excitation signal which, when fed into different filters, generates very accurate synthetic tones. Finally, a low-cost synthesis method is also proposed that can be used to generate very natural sounding percussive tones.**

I. Introduction

This work is concerned primarily with the analysis of percussive instrument sounds. In this paper, a musical instrument will be said to be a percussion instrument whenever the sound it produces results from the free vibration of a structure or a medium which has been set into motion by a short excitation. This designation includes classical percussion instruments (e.g, drums, vibraphone, marimba, bell, gong, piano, guitar etc.) and instruments that can be played as percussion instruments (e.g., pizzicati of violins, of cellos, or slaps of wind instruments). This paper will mainly present piano examples.
Percussion sounds are classically decomposed into two successive parts: the attack part, which is of short duration (less than a few hundred ms) and the resonance. The attack includes the interaction between the exciter and the resonating body, and lasts until the structure reaches a steady vibration. The resonance comes after the attack part, and includes the free vibration of the instrument body.

In theory, under the hypothesis of small perturbations, the resonance of the structure can be shown to generate a sound which is composed of exponentially damped sinusoids [1, 2]. The analysis of real sounds usually supports this theoretical result. The attack part of the signal however has a much more complex structure, and is usually quite difficult to analyze, to model, or to synthesize. This is unfortunate because it is well known that this very brief part is of extreme importance for the 'naturalness' and the recognizability of a musical sound. If the first few hundred milliseconds are suppressed from the highest notes of a piano, the resulting sounds lose their natural timbre and become almost unrecognizable. Very little is know about the nature of the attack part of percussive sounds, and this contribution is an attempt to help fill this gap.

This work focuses mainly on this very brief yet important segment of percussive sounds, the attack part. We have tried to model percussive sounds as the output of resonant filters excited by *short-duration* excitation signals. More specifically, the resonant filter represents the vibrating structure, while the short-duration excitation signal corresponds to the exciter. Our motivations are the following:

1. Better synthesis. While the filter concentrates all the information on the sinusoidal contents of the steady-state sound, the excitation signal can reproduce accurately the first hundred ms of the original signal, a dramatic improvement over classical synthesis techniques.

2. Deeper insight into the nature of the excitation. By extracting the short, very non-stationary attack part from the nearly-stationary sinusoidal components, it becomes possible to further analyze and model it.

3. Classification of percussion sounds on the basis of their excitation signal. Hopefully, the excitation corresponding to different notes played on the same instrument should keep some degree of similarity a fact which could be used to classify/recognize them.

The paper will be organized as follows: Different source/filter models are discussed in part II, and the calculation of the corresponding parameters is investigated part III. Part IV presents the application of such models to a musical instrument (the piano). Finally, we will review the advantages and drawbacks of our models, and discuss future investigations.

II. Models for percussive sounds

A. Why an Excitation/Filter Model?

Two major reasons why source/filter models seem particularly well suited for percussive sounds can be put forward.

These reasons are based on the physical properties of percussion instruments:

1. Sinusoidal contents: Under the hypothesis of small perturbations and linear elasticity, a freely vibrating body generates a sound which is composed of exponentially decaying sinusoids [1, 2]. Analyses of sounds of percussion instruments by classical techniques tend to confirm that after a certain time lag, percussive sounds can be fairly well approximated by sums of decaying sinusoidal components [3, 4].
   A simple way to generate decaying sinusoidal components consists in feeding a stable pole/zero filter with a short excitation signal. After the excitation has stopped, the output of the infinite impulse response filter is a sum of exponentially decaying sinusoids, provided the filter has no repeated pole [5].

2. Independence between the excitation on one hand, and the values of frequencies and damping factors on the other hand. The frequencies and damping factors of the decaying sinusoids observed in the resonating part of percussive sounds do not depend on the nature of the excitation: rather, they are determined by the physical properties of the resonating body, its dimensions, its stiffness etc. The excitation can only modify the relative amplitudes and phases of the sinusoidal components. This remark is true in theory as well as in practice. For example, the frequencies and exponential decays of the harmonics of a guitar sound *do not* depend on the way the string was set into vibrations. They are linked (among other things) to the tension, the length, the stiffness and the loss ratio of the material the string is made of, as well as to the string's boundary conditions [6]. The use of different excitation modes (plucking, hitting) only affects the relative amplitudes and phases of the harmonics.
   This property is also true for source/filter models: once the excitation has stopped, the frequencies and damping factors of the sinusoids composing the free response *do not depend on the excitation signal*, but only on the filter's poles.

It seems therefore quite natural to try to model percussive sounds as the output of an infinite impulse response filter (IIR) excited by a short-duration signal: The resonant filter generates the sinusoidal decaying part of the percussion signal, while the short-duration excitation reproduces its non-stationary onset.
From a physical point of view, this model can be interpreted as follows: the vibrating structure (e.g., the guitar strings and body, the vibraphone keys) is represented by the resonant filter, and the physical exciter (e.g., the guitarist's finger, the piano hammer, the vibraphonists mallet) is represented by the short-duration excitation signal. Much like the vibrating body of the real instrument, the resonant filter controls the frequencies and damping factors of the sinusoids in the quasi-stationary part. Much like the physical exciter, the short-duration excitation signal both controls their amplitudes and phases and accounts for the non-sinusoidal contents of the signal's onset.

B. Previous work in the area

In the past 20 years, source/filter modeling has been studied extensively, mainly in the domain of speech processing [7]. Applications to musical signals include works by Rodet &al. [8], Depalle [9], Potard &al. [10], Barrière &al. [11], although in the two last references, the source signal was not calculated, and assumed to be a simple impulse. By comparison to these works, our contribution focuses mainly on ways of obtaining the excitation signal. In addition, a multichannel excitation/filter model is presented which, to our knowledge, has never been applied before to musical signals.
The following section discusses a number of source/filter models for percussive sounds, and gives details on the nature and characteristics of both the resonant filter and the excitation signal.

C. Simple Excitation/Filter Model

C.1 Excitation signal characteristics

If the analogy 'physical-exciter $\leftrightarrow$ excitation-signal' is valid, we expect the latter to verify a number of properties:

1. The excitation signal should be of short duration. Because it represents the non-stationary part of the percussive sound, the excitation signal should not last much longer than this non-stationary part.

2. The excitation signal should not depend too much on the note played, but rather on the way the note was played. Within half an octave, the hammers of a piano can be considered similar from one note to another. The corresponding excitation signals should therefore exhibit similarities too. In particular, it should be possible to excite the filter corresponding to one note with the excitation corresponding to another note, a process called cross-synthesis [10].

3. Since the high frequency content of percussive instruments tends to increase when the excitation is harder or faster, the high-frequency contents of the excitation signal should increase in the same way.

Unfortunately, with the model used here, only point 1) is guaranteed. We will see that the excitation signals corresponding to different notes can actually appear quite different, and generate poor cross-syntheses.
Another possibility is to use a model where a single excitation signal is calculated to generate several sounds. This model is presented below.

C.2 Filter characteristics

The resonant filters used in this paper are classical rational transfer function filters: The Z transform $H(z)$ is of the form [5]:

$$H(z) = \frac{B(z)}{A(z)} \qquad \text{with} \tag{1}$$

$$B(z) = \prod_{i=1}^{q}(1 - w_i z^{-1}) \qquad \text{and}$$

$$A(z) = \prod_{i=1}^{p}(1 - z_i z^{-1})$$

where it can be seen that $A(z)$ and $B(z)$ are polynomials in the complex variable z^{-1}. The complex numbers w_i and z_i are, respectively, the zeros and poles of the transfer function $H(z)$. As mentioned above, the resonant filter accounts for the decaying sinusoidal part of the percussion signal. Since its zero-input response needs to decay to zero, the filter must be stable, or equivalently, its poles z_i need to lie inside the unit

circle $|z_i| < 1$ [5]. The frequencies f_i and damping factors α_i of the decaying sinusoids composing the zero-input response of filter $H(z)$ are given by:

$$\begin{aligned} f_i &= \frac{F_s}{2\pi}\arg(z_i) \qquad \text{and} \\ \alpha_i &= -F_s \log(|z_i|) \end{aligned} \tag{2}$$

in which F_s denotes the signal's sampling rate, $|z|$ and $\arg(z)$ denote respectively the modulus and the argument of complex z. We see that a slowly decaying sinusoid will correspond to a pole lying near the unit circle ($\alpha_i \approx 0$) while a highly damped sinusoid will be generated by a pole located closer to the origin. Eq. (2) shows that the values of the sinusoidal frequencies and damping factors depend only on the filter denominator $A(z)$, and conversely that $A(z)$ is completely determined by the values of frequencies and damping factors.

To preserve the analogy 'filter $\leftrightarrow$ instrument-body', we know that $H(z)$ should depend only on the resonance frequencies and damping factors in the resonating part of the sound, and *not* on their amplitudes and phases which are dependent on the excitation. Since the values of frequencies and damping factors are determined by filter $A(z)$ only, we see that *filter $B(z)$ can be chosen arbitrarily*, although this choice will later influence the excitation signal.

There are a number of possible choices for filter $B(z)$.

All–pole filter: If $B(z)$ is chosen equal to 1, the filter $H(z)$ is an all–pole filter:

$$\begin{aligned} H_{ap}(z) &= \frac{1}{A(z)} \qquad \text{or} \\ H_{ap}(z) &= \prod_{i=1}^{p} \frac{1}{1 - z_i z^{-1}} \end{aligned} \tag{3}$$

H_{ap} is a minimum-phase filter, i.e., all its zeros lie inside the unit circle (here, H_{ap} has no zero). This which means its inverse is a causal, stable filter a fact that will prove useful later. The filter H_{ap} may be realized a set of second–order sections in series. The typical transfer function of an all–pole filter is depicted in Fig. (1). Note the very deep high-frequency valley due to the choice of an all-pole form. We will see the importance of this remark later.

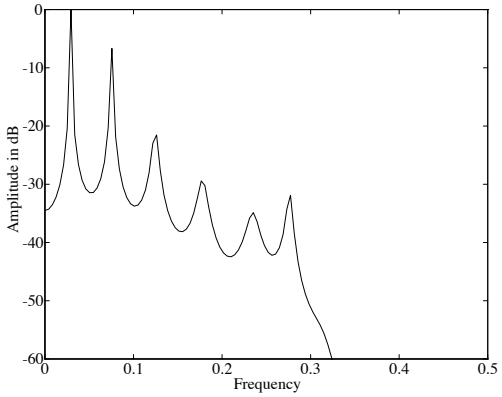

Figure 1: Magnitude transfer function of an all pole filter.

Parallel second-order sine sections: Another choice of filter $H(z)$ is the following:

$$H_{sc}(z) = \sum_{i=1}^{p/2} \frac{j}{1 - z_i z^{-1}} + \frac{-j}{1 - \overline{z}_i z^{-1}} \tag{4}$$

in which $\overline{z}$ denote the complex conjugate of z. In this form, the first order sections have been grouped into pairs of complex-conjugate poles. Filter H_{sc} is now composed of a set of real second–order sections in parallel. Its impulse response is a sum of exponentially decaying sinusoids with zero initial phase: a typical amplitude transfer function is depicted in Fig. (2). Note that unlike the previous all-pole form, H_{sc} exhibits zeros which interlace the poles near the unit circle. The zeros can be found by obtaining a common denominator in Eq. (4) and factoring the numerator.

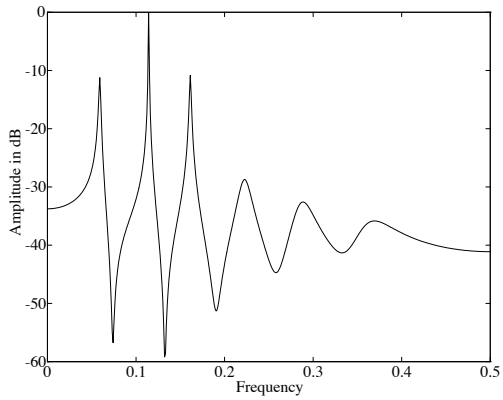

Figure 2: Magnitude transfer function of a filter built from second order sinusoidal sections.

Parallel second-order cosine sections: Finally, a third choice of filter $H(z)$ is the following:

$$\begin{aligned} H_{cc}(z) &= \sum_{i=1}^{p} \frac{1}{1 - z_i z^{-1}} \\ B(z) &= \sum_{i=1}^{p} \prod_{j \neq i} (1 - z_j z^{-1}) \end{aligned} \tag{5}$$

Polynomial $B(z)$ can be shown to be minimum phase (see appendix A) and consequently its inverse is a causal, stable filter. As in the previous case, H_{cc} is composed of a set of second–order sections in parallel, but this time its impulse response is a sum of exponentially decaying cosines, with zero phase. As shown by Fig. (3), the transfer function of such a filter typically exhibits much less deep valleys than both the serial implementation and the sum of sine sections. We will see in part B that this can be a very interesting property.

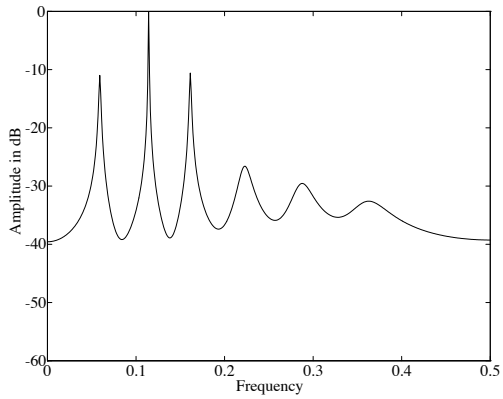

Figure 3: Magnitude transfer function of a filter built from second order cosine sections.

The three filters H_{ap}, H_{sc} and H_{cc} yield the same values of frequencies and damping factors, and differ only in the

sinusoidal phases and amplitudes in their impulse responses. Their respective characteristics will play an important role in the calculation of the excitation signals.

D. Single–Excitation/Multiple–Filter Model

In this part, we suppose we have a set of original sounds s_n^i corresponding to the same instrument, played in the same manner. For example, we might have an octave of piano notes played at the same velocity, or a set of vibraphone notes played with the same mallet at the same velocity.
If the double analogy 'physical-exciter ↔ excitation-signal', 'instrument resonator ↔ filter' is valid, and because the physical excitation remains about the same, there should exist *a single excitation signal* $e(n)$ and a set of resonant filters H_i such that all of the original sounds can be synthesized by feeding the same excitation signal into each of the resonant filter, as shown in Fig. (4). The validity of such a model is contingent on the

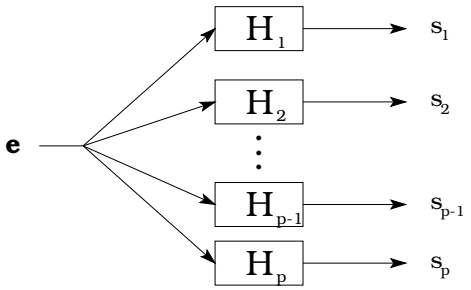

Figure 4: Single–excitation/Multiple–filter model. A single excitation e is fed into several filters H_i and generates several signals s_i.

fact that the physical excitation is the same for all signals. In practice, this can only be an approximation, but we will see that this approximation can be quite accurate. Obtaining sets of sounds corresponding to the same conditions of excitation is not always easy in practice. Part IV will discuss ways of achieving this.
The filters H_i used in this model are the same as in the preceding part $H_i(z) = B_i(z)/A_i(z)$. The filters $A_i(z)$ depend on the frequencies and damping factors of the sinusoids in the decaying part of signals s_i and the filters $B_i(z)$ can be chosen arbitrarily. It will be shown in the following sections that this model can help solve problems associated with the Simple Excitation/Filter Model.
The following part presents the calculation of the model parameters.

III. Estimation of Model parameters

Now that we have described our models for percussive sounds, we need to estimate their parameters: the filter coefficients, and the excitation signal.

A. Estimating the resonant filter

In this part, we suppose that we have an original percussive signal s that we try to model as the output of a resonant filter fed with a short-duration excitation signal. The estimation of the filter coefficients is the same for the single–excitation/single–filter model and for the single–excitation/multiple–filter model.
Since the filter numerator $B(z)$ is arbitrary, we need only estimate filter $A(z)$. Eq. (2) relates the zeros of polynomial $A(z)$ to the frequencies f_i and damping factors α_i of the sinusoids composing the sound's quasi-stationary part. The problem is therefore to estimate the set of f_i and α_i corresponding to the original percussive sound.

Various methods are available to estimate the frequencies and damping factors of sinusoidal components of a sound: parametric methods (Prony [3], MUSIC [12], ESPRIT [13]) and non parametric methods (Fourier transform, Wavelets, Time-Frequency analysis).

High–resolution parametric methods, like the Prony, MUSIC or ESPRIT methods, seem best adapted to the problem since they attempt to model the signal as the sum of exponentially damped sinusoids. Indeed, they are very good at detecting and estimating sinusoids in extremely short data-records. Unfortunately, they exhibit much worse performance with long-duration signals: in order to perform satisfactorily, they require the calculation of extremely large matrices (of the order of magnitude $N \times N$, where N is the total length of the signal). If N is not large enough, the estimated damping factors can exhibit serious bias [12]. Since, for a sampling rate of $32kHz$, the typical length of percussion sounds ranges from $100,000$ to $1,000,000$ samples, it becomes clear that the direct use of high–resolution methods is extremely impractical in our context. The problems associated with long duration signals can be worked around [14], but the procedure becomes lengthy and complex.

Non-parametric methods can also be used to estimate frequencies and damping factors. A classical way of estimating frequencies consists of calculating the power spectrum of a portion of the signal, performing a peak detection on the resulting spectrum, and fine-tuning the estimate of each frequency by use of parabolic interpolation [15]. The values of damping factors can be estimated by calculating the amplitudes of the sinusoids in two Fourier analyses located a different times. The damping factor is calculated from the rate of amplitude decay [10]. Unfortunately, because the sinusoidal decay is never perfectly exponential, this procedure yields estimates of damping factors which are highly sensitive to the placement of the two analysis windows. To circumvent this problem, we used a non-parametric analysis procedure inspired by Schroeder's method for the estimation of the impulse response of concert halls [16], and its time-frequency generalization ('Energy-Decay Relief') [17].

In the first step, successive power spectra $P_i(f)$ are calculated on a window located at increasing times n_i. These spectra are then accumulated, starting from the last one, finishing by the first one. This procedure yields a 3–D power spectrum $P_{cum}(f,i)$:

$$P_{cum}(f,i) = \log\left[\sum_{k\geq i} |P_k(f)|^2\right] \tag{6}$$

$$\text{with} \quad P_i(f) = \sum_{k=0}^{N-1} s_{k+n_i}.w_k.exp^{-j.2\pi.k.f/N}$$

where w_k represents a weighting window, usually a Hamming window. $P_{cum}(f,i)$ is the value at frequency f of the spectra accumulated over the range $k \geq i$.
In the second step, a peak detection is performed on the accumulated power spectrum $P_{cum}(f,1)$ to detect the sinusoidal components in the signal and calculate their frequency. Accurate frequencie estimates are obtained by use of a second–degree interpolation, much like in the standard procedure. Finally, the damping factors are estimated from the slope of the

3-D power spectrum, when f is kept constant and i is permitted to vary.
Because power spectra are non-negative, the backward-accumulation makes the 'Energy-Decay Relief' a non-increasing function of time for any given frequency. The corresponding 3-D plots are smoother than those obtained by the short-time Fourier transform. (See Fig. (10) in part IV.)

Once the values of the modal frequencies and damping factors have been estimated, the poles z_i are derived from Eq. (2). Finally, $A(z)$ is calculated from Eq. (1).

B. *Estimating the excitation signal: Simple Excitation/Filter Model*

We now turn to the problem of estimating the excitation signal. We will consider the two models (single-excitation/single-filter, single-excitation/multiple-filter) in two successive parts. Now that the filter denominator polynomial $A(z)$ has been estimated, and given a choice of numerator $B(z)$ the calculation of the excitation filter is a classical inverse filtering problem. This problem bears some similarities with the Linear Prediction Coding (LPC) inverse-filtering problem in which, given a speech signal and a vocal tract filter of the form $1/A(z)$, a glottal pulse excitation is searched that permits the best reconstruction of the original speech signal [7, 18]. However, the LPC filter and the corresponding excitation signals differ from ours: The LPC filter is generally smooth, with poles well inside the unit-circle, whereas our filters have highly resonant poles; The speech excitation signal is composed of a train of pulses embedded in noise, whereas our excitations typically have most of their energy concentrated in the first few hundred milliseconds
Two implementations can be used to compute the excitation signal: straight time domain inverse filtering, or frequency domain deconvolution.

B.1 Time domain inverse filtering

Inverse filtering consists of filtering the original signal by the inverse of $H(z)$, $H^{-1}(z) = A(z)/B(z)$. There are several problems associated with this technique:

Stability of $H^{-1}(z)$: If $B(z)$ has zeros outside the unit circle, $H^{-1}(z)$ is unstable and therefore, inverse filtering cannot be applied directly. One solution consists of splitting $B(z)$ into a minimum phase part $B_i(z)$ and a maximum phase part $B_o(z)$. $B_i(z)$ corresponds to the roots that are located *inside or on* the unit circle, and $B_o(z)$ corresponds to the roots that are located *outside* the unit circle. The signal s_n is filtered by $A(z)$, then by $1/B_i(z)$ and the result is time-reversed. Filter $1/B_o(z)$ is made minimum phase by reflecting its roots *inside* the unit circle, which amounts to reading its coefficients backward. This filter is used to process the time-reversed signal obtained in the preceding step. Finally, the result is time-reversed to yield the excitation signal e_n.
This process can be avoided by choosing a filter $H(z)$ that is minimum-phase, as is the case for the all-pole and parallel cosine-section filter structures.

Ill-posed nature of the inverse filtering problem: Depending on the filter $H(z)$, the problem of finding the excitation signal corresponding to the original signal can be *ill-posed* [19]. This means *many different excitation signals can yield quite similar results* (in the sense of the L^2 norm), after filtering by $H(z)$. Equivalently, the excitation signals e_n corresponding to two very similar signals x_n can be extremely different a fact that is highly undesirable.
Here is an illustration of this problem. Suppose $H(z)$ is an all-pole filter with the amplitude transfer function described in Fig (5). A white noise (the excitation signal) w_n is fed

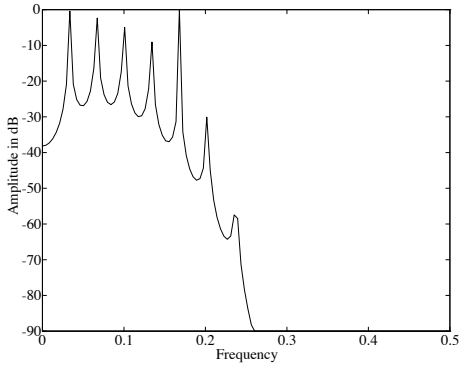

Figure 5: Magnitude transfer function of a resonant all-pole filter.

into filter $H(z)$ and a second, independent white noise b_n is added to the result at a $-70dB$ level, yielding what we will call the original signal x_n. We have $x_n = h_n \star w_n + b_n$ in which h_n is the filter's impulse response, and b_n can represent quantization or measurement noise. The Fourier spectrum of the signal x_n is given in Fig (6). The result of time domain inverse filtering is given in Fig (7) in the frequency domain.

Although the original signal x_n had almost no energy in

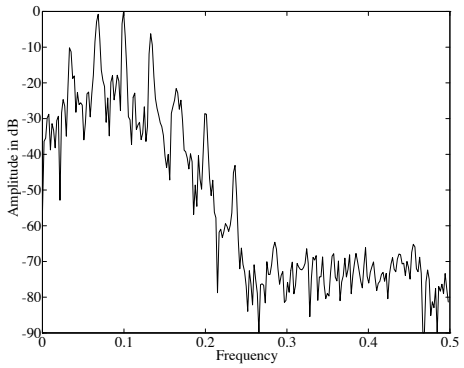

Figure 6: Fourier analysis of the original signal.

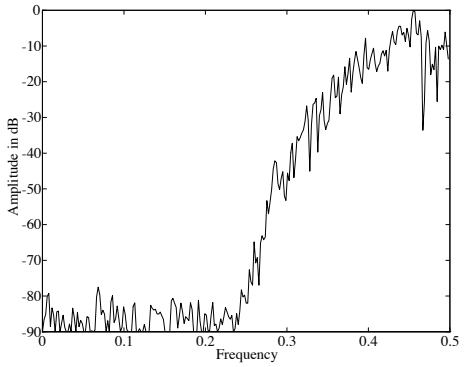

Figure 7: Fourier analysis of the excitation obtained by time domain inverse filtering.

the high-frequency domain, the calculated excitation signal e_n

is submerged by high-frequency components. However, these high-frequency components are *irrelevant* because they have almost *no influence* on the reconstructed signal $\hat{x}_n = h_m \star e_m$ because the transfer function of $H(z)$ is extremely attenuating in the high-frequency region. We see that high-frequency components are almost completely filtered-out by $H(z)$In fact, one can remove the frequency content of the calculated excitation signal above $f = 0.35$ and still obtain a reconstructed signal that is *indistinguishable* from the original signal.

The ill-posed nature of inverse filtering has been studied extensively [19]. Our problem is that a more 'regular' solution which is not quite accurate would be more useful to us than the highly irregular, exact solution. In the case shown above, 'regular' means 'with little high-frequency content'. The ill-conditioning problem can be worked around in several ways.

1. Wiener Filtering: This technique, used in signal enhancement [20] can be applied to inverse filtering. It requires the a-priori knowledge of the power spectral densities $P_{bb}(f)$ and $P_{ww}(f)$ of the corrupting noise and of the excitation signal respectively [21]. In our problem, we would need to estimate both.

2. Regularization techniques: The idea here is that the exact solution is no longer sought, but instead, a nearly exact, regular solution is computed [19]. Regularization techniques make use of a-priori knowledge of the excitation signal itself, and also require fine tuning of the so-called smoothing parameter to give good results [22]. Regularization techniques and Wiener filtering are closely related and yield the same type of equations for the calculation of the excitation signal.

3. The Moore-Penrose pseudo-inverse techniques, based on the generalized inverse of a singular matrix, can also be applied on inverse-filtering problems [23]. However, pseudo-inverse methods require the estimation of the null-subspace of the filtering matrix (which is not a simple task), and more importantly they prove impractical for long duration signals, because of the size of the matrices involved.

Other techniques can also be used, all of which impose a constraint on the sought solution.
Another simpler way to avoid this problem consists of using filters $H(z)$ which are better conditioned. For inverse filtering problems, ill-conditioning can be measured by the ratio of the maximum to the minimum of the filter's amplitude transfer function [24]. The higher the ratio, the worse the conditioning of the problem. The filter of Fig (5) is obviously very ill-conditioned. Most of the time, the all-pole implementation turns out to be highly ill-conditioned, especially if there is no pole in the high-frequency region. The 'sum of sinusoidal sections structure' also exhibits deep valleys as can be seen in Fig (2), and therefore it tends to be ill-conditioned. In contrast, the structure 'sum of cosine sections' is generally better-conditioned because its amplitude transfer function does not exhibit highly attenuated regions as can be seen in Fig (3). This fact favors its use in our context. We will see some examples of this remark later on.
Because regularization techniques such as the ones described above are difficult to use in practice (sensitive parameter-tuning, a-priori information not readily available), we decided to use well-conditioned filters for which straight inverse-filtering gives good results.

Poor conditioning of filter coefficients: $A(z)$ is determined by the values of its roots; expressing $A(z)$ as a polynomial can lead to extremely large coefficients, especially if the roots are close to each other. In such a case, when n first order sections are multiplied together, the middle coefficients of the resulting polynomial can reach values as large as $\frac{n!}{[(n/2)!]^2}$, with first and last coefficients near unity. This wide coefficient range is likely to generate numerical instabilities in the process of inverse filtering as shown below.
Any numerical error in the polynomial coefficients a_k causes a displacement of the polynomial root z_i given by [5]

$$\frac{\partial z_i}{\partial a_k} = \frac{z_i^{N-k}}{\prod_{l=1, l\neq i}^{N}(z_i - z_l)} \tag{7}$$

in which N is the number of roots. When the roots are tightly clustered, two factors combine their effects: the denominator of Eq. (7) tends to be large, making the root locations very sensitive to small errors in the coefficients a_k, and the a_k tend to span a large dynamic range a fact that favors rounding errors. For example, consider an harmonic, non-decaying signal containing 20 harmonics with fundamental frequency $100Hz$ sampled at $48kHz$. A direct application of Eq. (7) shows that the sensitivity of the 10th root with respect to any polynomial coefficient a_k is (in modulus) larger than 10^{35}. In other words, to obtain a 1% precision on the root location, we would need to have a precision of 10^{-37} on the filter coefficients, which is unattainable in practice.
In the general case, finding a solution to this problem is not easy. However, if $H(z)$ is an all-pole filter, then $H^{-1}(z)$ can be split into several sub-filters with interleaved zeros. Each sub-filter has fewer zeros which, in addition are more distant from each other. Consequently, the zeros of the sub-filters are better conditioned than those of the original inverse filter $H^{-1}(z)$. Inverse filtering is carried out by successively applying each sub-filter.
When $H(z)$ is a sum of second-order cosine or sine sections, $A(z)$ can still be decomposed into several sub-filters, but the numerator $B(z)$ cannot be explicitly expressed as a product of sub-factors because its roots are not readily available (finding the roots of $B(z)$ would require finding $B(z)$ from the second-order sections and factoring, an ill-conditioned process). In this case, the decomposition into successive sub-filtering processes is not possible, and an alternative to time domain inverse filtering must be used. Note that down-sampling the original signal usually remedies the coefficients conditioning problem by increasing the distance between the polynomial roots.

When to use time domain inverse filtering? Time domain inverse filtering can be used whenever the resonant filter is minimum-phase, is well-conditioned, and does not exhibit large coefficients. We found this to be the case for many signals when the sum-of-cosines structure was used.
In the following part, we present an alternative to time domain inverse filtering.

B.2 Frequency domain deconvolution

With certain precautions, inverse filtering can be performed in the frequency domain by use of the Discrete Fourier-Transform (DFT). It is well known that the circular convolution of two signals u_n and v_n can be performed by multiplying their complex DFTs U_k and V_k and taking the Inverse Fourier Transform [5]. Of course, the major advantage of such a scheme is the possibility of using Fast Fourier Transform. A sufficient condition for the circular convolution to be equivalent to an

acyclic convolution is that the DFT length N should satisfy

$$N > N_u + N_v - 1 \tag{8}$$

where N_u and N_v denote the lengths of the two signals u_n and v_n respectively. In other words, the size N of the DFT should be larger than the length of the signal $u \star v$, where $\star$ denotes acyclic convolution.

It follows that deconvolution can be performed in the frequency domain by use of the Discrete Fourier Transform: the DFTs of the original signal x_n and of the impulse response h_n of filter $H(z)$ are calculated, making sure the size N of the DFT satisfies the condition stated by Eq. (8). For each discrete frequency, the former is divided by the latter, and an inverse Fourier Transform of the result is taken, yielding the excitation signal.

The condition stated by Eq. (8) specifies that the length N of the DFT should be larger than the length of the convolved signal, x_n (since by hypothesis, x_n is the convolution of h_n by e_n). When this condition is met, the frequency domain deconvolution is absolutely equivalent to time domain inverse filtering by the (possibly non-causal) stable filter with transfer function $H(z)$.

In theory, since filter $H(z)$ generally has an infinite impulse response h_n, the above condition is not fulfilled and time aliasing is likely to occur. However, if N is large enough, the impulse response h_n can be considered null outside the range $-N < n < N$.

The practical implementation of frequency domain deconvolution suffers less limitations than time domain inverse filtering, as will now be shown.

Stability of $H^{-1}(z)$. The frequency domain deconvolution can be performed even if filter $H(z)$ is not minimum-phase. The inverse filter $H^{-1}(z)$ is used in its (possibly non-causal) stable version (its impulse response can extend toward $\pm\infty$). This avoids the double filtering scheme that was required by time domain inverse filtering.

Ill-conditioning. Because ill-conditioning is intrinsic to our problem, the remarks we made for time domain inverse filtering also apply for frequency domain deconvolution.

Filter-coefficient stability-problem. The filter's recursive form coefficients are no longer necessary. Only the impulse response is needed, which can easily be calculated even when the filter includes a large number of sinusoidal components. If necessary, the transfer function is expanded as a sum of second-order sections and the impulse responses of all sections are added.

In summary, for the single-excitation/single-filter model, the most difficult problems encountered during the calculation of the excitation signal are those related to the ill-conditioned nature of inverse filtering processes. When a filter is used that exhibits an amplitude transfer function with a large dynamic range, the excitation signal is often plagued with unwanted artifacts, e.g., high frequency noise, sinusoidal resonances etc. In practice, we have found that only the sum of cosine sections gave good results, because of its generally well-behaved transfer function.

C. Estimating the excitation signal: Single-Excitation Multiple-Filter Model

C.1 The least-squares deconvolution

When the single-excitation/multiple-filter model is used, the problem of finding an excitation signal corresponding to several pairs of original-sounds/resonant-filters no longer admits an exact solution since there are more equations than there are unknowns. In such cases, it is natural to turn to a least-squares solution. The problem can be stated as follows: Given p original signals x_n^i $(0 \leq i < p)$ and p resonant filters H_i with impulse responses h_n^i we search for an excitation signal e_n which minimizes the cumulative error $\mathcal{E}$ defined as:

$$\begin{aligned} \mathcal{E} &= \sum_{i=0}^{p-1} \left(\sum_{k=0}^{N-1} (x_k^i - \hat{x}_k^i)^2 \right) \qquad \text{with} \\ \hat{x}_k^i &= h_n^i \star e_n \end{aligned} \tag{9}$$

where $u \star v$ denotes the convolution of the two signals. Eq. (9) states that the cumulative error $\mathcal{E}$ is the sum of the quadratic errors between the original signals x_n^i and their synthetic versions $\hat{x}_k^i$ (obtained by feeding e_n through each resonant filter H_i).

An important point is that the minimum $\mathcal{E}_{min}$ of the cumulative error $\mathcal{E}$ is a measure of the adequacy of our model: if $\mathcal{E}_{min}$ is found to be large (in a sense that needs to be specified), then the single-excitation/multiple-filter model can be judged inadequate. On the contrary, a small value of $\mathcal{E}_{min}$ tends to validate the model. This is one important advantage of the single-excitation/multiple-filter model over the single-excitation/single-filter model.

It is easily shown (see appendix B) that the least-squares solution can be expressed in the frequency domain by:

$$E(f) = \frac{\sum_{i=0}^{p-1} H_i^*(f) X_i(f)}{\sum_{i=0}^{p-1} H_i^*(f) H_i(f)} \tag{10}$$

in which capital letters refer to Fourier Transforms of the corresponding signals. Thus $E(f)$ is the Fourier Transform of the excitation signal, $H_i(f)$ is the complex transfer function of filter H_i, and $X_i(f)$ is the Fourier Transform of original signal x_n^i. The time domain excitation signal is finally obtained by an inverse Fourier Transform.

The Discrete Fourier Transform can also be used here, under conditions similar to those stated in Eq. (8). More precisely the length N of the discrete Fourier Transform needs to satisfy

$$N > N_x^i \qquad 0 \leq i < p \tag{11}$$

in which N_x^i represents the length of the original signal x_n^i. Strictly speaking, none of the x_i is of limited duration (since they are the outputs of IIR filters H_i fed by a time-limited excitation). However, provided the size N of the FFT is large enough, they can be considered to be of finite duration to a good degree of approximation. Of course, the case $p = 1$ corresponds to the frequency domain calculation of the excitation signal in the single-excitation/single-filter context.

As will now be shown, some of the problems encountered for the single-excitation/single-filter model no longer exist in the present context, and new problems arise.

C.2 Problems and solutions

Stability of the inverse filters H_i^{-1}. Since the calculation of the excitation signal is performed in the frequency domain, the same remarks apply as in paragraph (B.2): when the filters

H_i are not minimum phase, deconvolution is achieved using the non-causal, stable form of their inverses.

Ill-posed nature of the inverse filtering problem. Eq. (10) with $p = 1$ shows that whenever $H(z)$ has a frequency region of low energy ($|H(f)| \approx 0$), the ratio tends to grow toward ∞, which illustrates the ill-conditioning of the inverse filtering problem. We see that using different filters H_i tends to alleviate this problem by minimizing the chances that the denominator in Eq. (10) be zero. In fact, if the respective regions where the filters H_i have a low amplitude-response do not overlap, the sum of the filters' amplitude transfer functions is never near zero, and the problem becomes well-conditioned.
As a result, in the single-excitation/multiple-filter context, the inverse filtering problem appears better conditioned. An intuitive way to understand why the problem of determining the excitation should be better conditioned is to remember that ill-conditioning results from the existence of many very different 'near-solutions' (see section (B)). In the single-excitation/multiple-filter context, the problem is more selective because we impose more constraints on the desired solution (to the point that there no longer exists an exact solution). As a result, it seems natural that there should be fewer suitable 'near-solutions'. In fact, many techniques for regularizing ill-conditioned problems are based on the idea of imposing additional constraints on the solution (e.g., regularity criteria) [19].

Filter-coefficient stability-problem. As mentioned in the single-excitation/single-filter context, working in the frequency domain eliminates the problems associated with the filter coefficients. This remark remains valid in the single-excitation/multiple-filter context.

The synchronization problem. The single-excitation/multiple-filter model can be valid only if the original signals x^i were recorded simultaneously. To see that, consider original signals whose onsets occur at non-simultaneous times: since the group-delays of the filters H_i cannot be adjusted, our chances to find a suitable common excitation are weak. Indeed, when such a case occurs, the calculated excitation-signal is smeared and looses its percussiveness: the cumulative error of Eq. (9) tends to be large and the quality of the syntheses is dramatically altered.
In most cases, the original signals do not include an absolute time-reference, and they need to be synchronized in some manner before the common excitation can be calculated.

One way to achieve this consists in calculating the excitation signals e_n^i corresponding to each pair of resonant-filter H_i / original-signal x_n^i (momentarily using the single-excitation/single-filter model). Either of the methods suggested in section (B) can be used. Each excitation signal e_n^i reflects the onset of the corresponding original signal and incorporates the group-delay of the corresponding resonant filter. It is now possible to mutually synchronize the excitation signals by use of a simple, classical cross-correlation technique: the time domain cross-correlation $r_{i,j}(\tau)$ of two excitation signals e_n^i and e_n^j is calculated, and the value of τ that corresponds to its maximum gives an estimate of the delay $\tau_{i,j}$ between the two signals.

$$\begin{aligned} r_{i,j}(\tau) &= \sum_{n=1}^{N} e_n^j e_{n+\tau}^i \\ \tau_{i,j} &= \tau \leftrightarrow \max_{\tau} r_{i,j}(\tau) \end{aligned} \tag{12}$$

This procedure makes it possible to mutually synchronize the excitation signals, and therefore the original signals. After synchronization, the common excitation can be calculated.
To avoid the costly calculation of each excitation signal, one could be tempted to apply the cross-correlation technique to the original signals rather than to the excitation signals. There are two major reasons for not doing so:

1. The original signals are usually very different from one another both in the time domain and in the frequency domain whereas the excitation signals exhibit many similarities (see example below). The cross-correlation technique only works when the signals are similar.

2. The cross-correlation technique works best when the signals are close to white noise because their cross-correlation exhibits a large peak at $\tau = \tau_{i,j}$ and few secondary peaks. On the contrary, when the signals are near-periodic, the cross-correlation exhibits a number of peaks at multiples of the periods, and it is rarely clear which one is the main peak.
   In our case, even when the original signals exhibit harmonic structures, the excitation signals are free of periodicity, which makes them easier to synchronize.

The following example presents the synchronization of four piano notes. Fig. (8) presents the four original signals, and Fig. (9) the four excitation signals. The filters were of the parallel cosine-sections form and the excitations were calculated by use of the frequency domain deconvolution technique. While the original signals seem very different, the excitation signals exhibit very similar time domain shapes, a fact which makes their synchronization easier. It is quite clear that the two signals at the bottom of Fig. (9) are nearly synchronized, while the two top ones need to be time-adjusted. Indeed the cross-correlation technique confirms this visual clue.

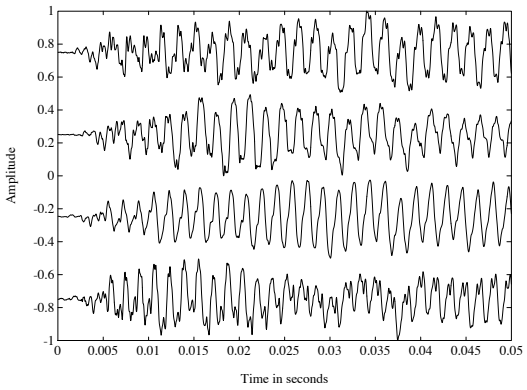

Figure 8: Time domain representation of the first 50ms of four piano notes.

The necessity of rescaling. If the original signals were not recorded at exactly the same level, amplitude discrepancies can be observed between different specific-excitations. In this case, a rescaling stage is needed to adjust the amplitudes of the original signals so that the specific excitations are of comparable energy. This can be done at the same time the synchronization is performed: the energy of each specific excitation is calculated and the corresponding original signal rescaled accordingly. This rescaling stage solves the amplitude discrepancy problems that can occur when the original sounds are recorded during different sessions.

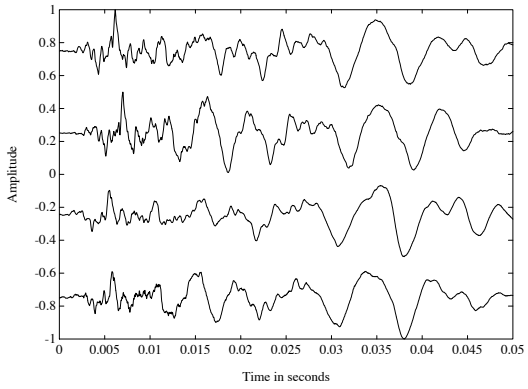

Figure 9: Time domain representation of the excitations corresponding to four piano notes. Note the great similarity between the signals.

Now that the general framework has been described, we will demonstrate how the different models apply to the sound of a real piano.

IV. Application to real signals

In all that follows, we will consider both the single-excitation/single-filter and the single-excitation/multiple-filter models. The excitation signals corresponding to the first model will be called 'specific excitations' and those corresponding to the second model will be called 'common excitations'.

The piano was chosen because of its very specific onset. For high-frequency notes (e.g., above C3), the percussive sound generated by the hammer when it hits the string is quite audible, and has been found to be very important for the recognizability and the naturalness of the sound [25, 26]. Moreover, the onset is very difficult to synthesize by classical means (additive synthesis, frequency modulation etc...) to the point that synthesizer manufacturers often resort to hybrid techniques in which the (previously sampled) onset is simply played back, and only the resonance is synthesized.
In this part, we will describe how the single-excitation/multiple-filter model can be applied to piano sounds. We first turn to the problem of obtaining valid original signals.

A. Selection of the original piano sounds.

As mentioned in section D, for the single-excitation/multiple-filter model to be valid, we would like the physical-excitation parameters to be uniform across all considered notes. Two of these parameters seem of the highest importance: the size and weight of the hammer, and its velocity.
In the case of the piano, the condition of uniformity cannot be perfectly met since the hammers used on different notes are not the same. However, if we restrict our choice to notes belonging to the same octave or half-octave, the characteristics of the felt remain the same and the corresponding hammers keep about the same size and weight and can be considered to excite the strings in a very similar way.
More difficult is the problem of the hammer velocity. it is well known that the velocity of the hammer not only controls the loudness of the sound but also influences its spectral contents in a non-linear way [27]. Consequently, it is very important to be able to control the velocity of the hammer corresponding to the recorded sound in order to make sure the excitation is uniform. This is not very easy in practice but there are a number of ways of achieving this task:

- Have a professional pianist play the notes at the same nuance (piano, mezzo, forte etc...)
- Design a special device that hits the piano keys with a constant force/speed.
- Send notes with a constant velocity parameter to a midi-driven piano (e.g., Yamaha's synclavier).

We used both first and third methods, with similar results. The examples shown below were simply extracted from a sample CD (Mc Gills University Master Samples), using series corresponding to the nuance 'piano'. We used a series of 7 semitones, $C5$, $C5^\sharp$, $D5$, $D5^\sharp$, $E5$, $F5$ and $F5^\sharp$, with respective fundamental frequencies $523.24Hz$, $554.36Hz$, $587.32Hz$, $622.26Hz$, $659.26Hz$, $698.46Hz$, $739.98Hz$,.

B. Analysis/synthesis.

Calculation of the resonant filters. The resonant filters corresponding to each notes were determined by use of the cumulated Fourier Transform method described in section A. Fig. (10) presents a three-dimensional representation of the cumulated spectrum corresponding to the note $D5$.

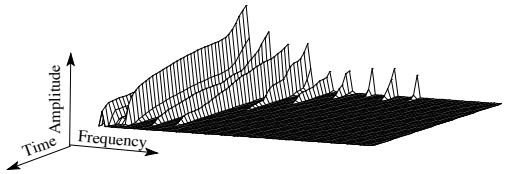

Figure 10: Three-dimensional representation of the cumulated spectrum corresponding to the note $D5$.

Synchronization: calculating the specific excitations. As mentioned in section C, the synchronization of the original signals need to be checked and adjusted if necessary. This requires the calculation of the specific excitations corresponding to our 7 notes. Fig. (11) presents the specific excitations corresponding to the note $D5$ when the parallel sine-cell filter (top signal) and the parallel cosine-cell filter (bottom signal) are used. The top signal is plagued with extraneous high-frequency components that hide most of its time-domain features. The bottom signal gives much more detail on the nature of the excitation. Note however that both signals allow a perfect reconstruction of the original piano sound, when fed through their corresponding filters.
The cross-correlation technique was applied to all seven specific excitations to mutually synchronize the original piano sounds.

Calculation of the common excitation. Eq. (10) suggests that the common excitation can be computed iteratively. The contribution of each new signal is added to the numerator and to the denominator, and the common excitation is updated. Thus, when $i = 0$ we have the specific excitation corresponding to signal x_n^0, when $i = 1$ we get the excitation common to signals x_n^0 and x_n^1, etc... Finally, $i = 6$ gives us the common excitation corresponding to the seven original piano sounds. Fig. (12) presents the excitations obtained successively for $i = 0$, $i = 2$, $i = 4$ and $i = 6$. We can remark that the time-domain representation of the excitation signal changes only slightly as the excitation becomes common to more and more signals.

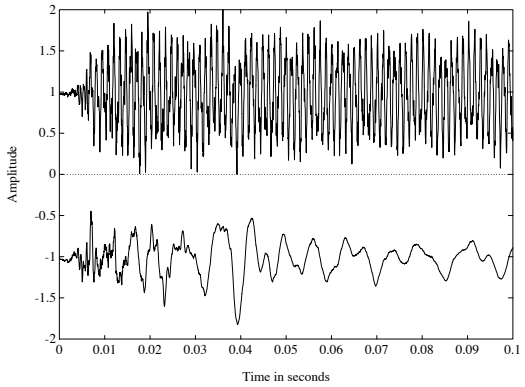

Figure 11: Specific excitations of the note $E5$. The excitations were obtained by frequency-domain deconvolution. The top signal corresponds to the parallel sine-cell filter, and the bottom signal to the parallel cosine-cell filter.

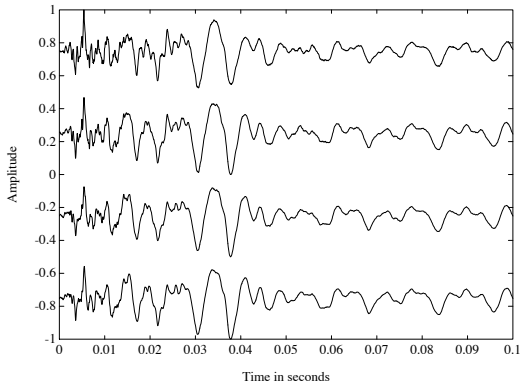

Figure 12: Successive common excitation signals corresponding to (from top to bottom) one piano note, three notes, five notes and seven notes.

C. Results and discussion.

C.1 Common excitation signal.

A careful inspection of the common excitation signal leads to the following remarks:

Time-domain aspect: Most of the energy in the common excitation is lumped in the first $400ms$ as can be seen on Fig. (13). The resulting sound is a loud, short-duration percussive burst which appears quite similar to the sound of a piano whose strings are not allowed to vibrate (e.g., muted by a thick felt). The common excitation includes the low-frequency vibration of the soundboard which was not accounted for by the resonant filters. The resonance of the soundboard although much more damped than the vibration of the strings, is responsible for the non-zero tail of the excitation signal.

Frequency-domain aspect: Fig. (14) presents the spectrum of the first $128ms$ of the excitation signal. One remarks the strong low-frequency contents, and a number of weak spectral rays which demonstrate the existence of remaining low-amplitude sinusoidal components. The low-frequency contents corresponds to the vibration of the soundboard, and accounts for the most part of the excitation signal. The frequencies of the low-amplitude sinusoids correspond to the frequencies of the sinusoids present in the original piano tones. We will see in the following part that these sinusoidal components account for the beatings in the original piano tones.

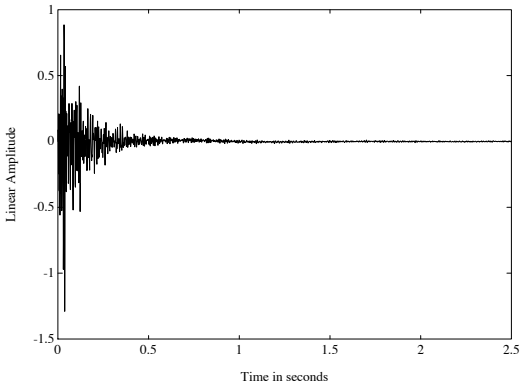

Figure 13: Time domain representation of the excitation signal common to seven piano notes.

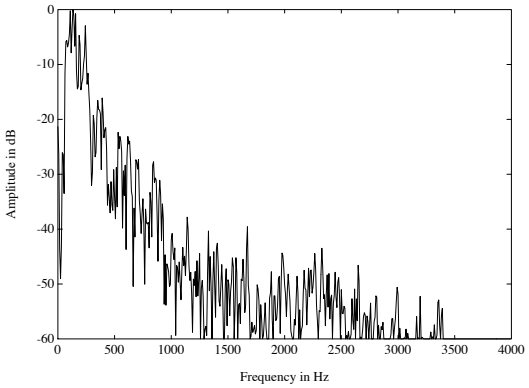

Figure 14: Spectrum of the common excitation signal. The window size is 128 ms, and a hanning weighting was applied to the signal.

C.2 Synthetic piano tones.

Fig. (15) and Fig. (16) give the three-dimensional representation of one of the original signals (note $F5$, $698.46Hz$) and of its synthesis obtained by feeding the common excitation signal into the resonant filter corresponding to the note $F5$.

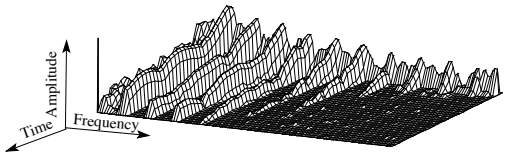

Figure 15: Three-dimensional representation of an original piano tone. The note is $F5$. The fourier window size was 1024pts, the increment was 2048pts and a Blackman weighting window was used. For readability, only the frequency band between 0 and $8kHz$ is shown.

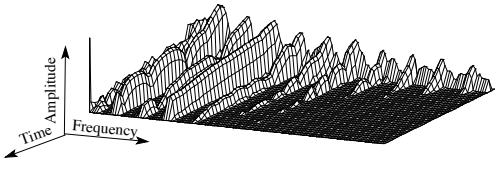

Figure 16: Three-dimensional representation of a synthetic piano tone. The note is $F5$. The fourier window size was 1024pts, the increment was 2048pts and a Blackman weighting window was used. The synthesis was obtained by feeding the common excitation signal into the resonant filter corresponding to the note $F5$.

Transient response: The transient response is well reproduced. The first milliseconds of the synthetic signal bear a great similarity with those of the original signal. This visual clue is confirmed by listening tests. The attack of both sounds are similar to the point that it becomes very difficult to differentiate them. In particular, the sound of the hammer, and the very damped resonance of the soundboard are present in all synthetic sounds. This, of course, is an advantage inherent to any source-filter model. Our additional improvement comes from the fact that only one excitation is used for seven piano sounds, with very satisfying results.

Double exponential-decay and component beatings. Piano sounds exhibit two specificities that are very important for their naturalness, the double exponential-decay phenomenon and the amplitude beatings.

The double exponential-decay is visible in Fig. (10). In a dB scale, the amplitude of each sinusoid is composed of two straight lines with different slopes. This fact is well known and has been studied extensively in [28]: due to coupling between orthogonal polarizations of vibration of the string, the exponential-decay of each component shifts to a lower value after a short time-lag. The second, less damped vibration gives the piano its long sustain.

In our source-filter models, there are two ways of accounting for this double exponential-decay: the double decay can be included in the filter structure or coded by the excitation signal.

Including the double exponential-decay into the filter structure can be done by allocating two pairs of complex poles to each sinusoidal component. One pair accounts for the first damping factor, and the other pair for the second, lower damping factor, according to Eq. (2). Unfortunately, because the two pairs of poles are extremely close in the complex plane, this solution leads to filter-coefficient instabilities, and poor filter-conditioning.

The second option consists in letting the excitation signal handle the double exponential-decay. This is the solution we chose. The question remains to decide which decay should be included in the resonant filter. In order to make the excitation signal as short as possible (i.e. with all its energy concentrated in the first milliseconds), the smallest exponential-decay was assigned to each sinusoidal component. On the contrary, assigning the largest decay would cause the excitation to contain more energy in the sustain part of the piano sound. A comparison between Fig. (15) and Fig. (16) demonstrates that this solution is quite efficient: the double exponential-decays are similar in both sounds (observe the higher frequencies).

The amplitude beatings give the piano sound its naturalness. Weinreich in [28] has shown that amplitude beatings result from the coupling of adjacent strings with slight mistuning (most of the piano notes correspond to pairs or triplets of strings). The result is a slow time-evolution of the timbre and amplitude of the piano tone. Here again, there are two ways of accounting for this phenomenon in the context of source-filter models. The beatings can be included in the resonant filter, or left in the excitation source.

Beatings can be generated by assigning two or three pairs of poles for each sinusoidal component, with slightly different frequencies (Eq. (2)). As a result, each harmonic component is modeled by a sum of two or three sinusoids with close frequencies: provided the frequencies are close enough, the resulting sound exhibits low-frequency beatings similar to those observed in original piano sounds. This solution has the drawback of making the resonant filter two to three times as complex.

When the resonant filter is assigned only one pair of poles per sinusoidal component, the amplitude beatings are coded in the excitation signal. The excitation signal's energy is still concentrated in the first milliseconds, but the excitation also includes sinusoidal components which serve to control the beatings. A careful audition of the excitation signal reveals the presence of a weak sound which resembles to a mix of all seven original signals. Indeed, if the excitation is truncated after a few milliseconds, the transient part of the piano tone is preserved, but the beatings are eliminated and the resonance presents a very unnatural steadiness.

The results we described for the note $F5$ are indeed valid for all seven notes. The synthetic tones are characterized by a remarkable naturalness both in their transient part and in their resonance: the impact of the hammer and the very low-frequency resonance of the soundboard are well reproduced in all synthetic sounds. The beatings are not always exactly similar to those of the original sounds, but sound quite realistic (in fact, much more natural than the generally too-regular beatings obtained by pairs or triplets of close sinusoids). The cumulated error $\mathcal{E}$ is $-15.32dB$ below the average rms energy of the original sounds. Although low rms-error cannot be directly interpreted as auditory similarity (for example, the rms-error is sensitive to the phases of the components) this result shows a good fit of the model to the data.

However, the harmonics of some of the synthetic sounds have amplitudes that can be very different than those of the corresponding original sounds. We will now analyse this problem and suggest ways of solving it.

D. The problem of overlapping frequencies.

A careful inspection of Fig. (15) and Fig. (16) shows that the third harmonic is assigned a larger amplitude in the synthetic sound than in the original sound. A listening test reveals that the two sounds, although perceptually very similar, have a slightly different timbre. To understand where this problem comes from, we need to gain better insight into the single-excitation/multiple-filter model.

In any source/filter model, the excitation has two main functions: 1) it controls the amplitudes and phases of the sinusoidal components, 2) it reproduces the non-sinusoidal part in the original signal. Let us take a closer look at the first point. The control of the sinusoidal amplitudes is achieved by adjusting the spectral contents of the excitation in the frequency areas located around the sinusoidal components. This is because the amplitudes A of the sinusoid of frequency f_i in the synthetic sound are given by

$$A = |H(f_i)| \times |E(f_i)| \tag{13}$$

where $H(f)$ and $E(f)$ are the Fourier transforms of the filter's impulse response and of the excitation signal.

In the single-excitation/multiple-filter model, the excitation needs to control the values of the sinusoidal amplitudes *for each sinusoidal component of each synthetic signal*. Suppose a sinusoid of one of the original sound x_n^i has a frequency that is very close to that of a sinusoid in another original tone x_n^j (this can be the case for two harmonic signals whose fundamental frequencies are rationally related). Then Eq. (13) must be true for both signals:

$$\begin{cases} A^i &= |H_i(f)| \times |E(f)| \\ A^j &= |H_j(f)| \times |E(f)| \end{cases} \tag{14}$$

Obviously, this can be exactly satisfied only if

$$\frac{A^i}{|H_i(f)|} = \frac{A^j}{|H_j(f)|} \tag{15}$$

a condition that is not necessarily met. In other words, if the harmonics of two original signals have nearly equal frequencies, then the excitation cannot independently adjust their respective amplitudes in the synthetic signals, and some spectral distortion occurs.
We can suggest two solutions to this problem:

Iterative fine tuning of the model. One way to overcome this problem, and achieve a better set of synthetic sounds consists in adjusting the numerators of filters H_i: bear in mind that the only constraint on the resonant filters concerned the location of their poles (part. C). The numerators B_i were arbitrarily chosen. Now these additional degrees of liberty can be used to carefully adjust the amplitudes of the sinusoids in the synthetic sounds. A simple way to achieve this consists in replacing the unity residues in Eq. (5) by non-unity values, leading to a new formula for the modified resonant filters H'_i:

$$H'_i(z) = \sum_{j=1}^{p} \frac{r_j}{1 - z_j z^{-1}}$$

The residues r_i are calculated from the difference between the amplitude A of the sinusoid in the original signal x_n^i and the amplitude $\hat{A}$ of the same sinusoid in the synthetic signal $\hat{x}_n^i$.
This slight modification helps reduce the amplitude-errors mentioned above, and can improve the resemblance between original sounds and synthetic sounds.
To gain additional quality, we can remark that the common excitation calculated for the resonant filters H_i is no longer necessarily the solution of the minimization problem stated in Eq. (9) when the filters H_i are replaced by H'_i. The search for an optimal solution both for the filters and the excitation signal then leads to an iterative bi-linear procedure in which the common excitation is computed, the filters' residues modified according to the contents of the synthetic sounds, the excitation re-computed and so on. Each step of the iterative procedure is a least-squares minimization, and therefore is guaranteed to diminish the preceding value of the error $\mathcal{E}$ defined in Eq. (9). The series of the successive errors $\mathcal{E}_l$ is thus guaranteed to decrease monotonously.
In practice, this iterative procedure reduces only by a few dB the rms error between the original sounds and their respective syntheses. The improvement of the sound quality is sometimes difficult to detect by ear. This negative result comes from the fact that we force each sinusoidal component to have the *same overall rms-power* in both signals. A problem remains however, when the sinusoidal amplitudes have *different time-variations*: if the harmonics of two original signals have nearly equal frequencies, then the common excitation cannot independently adjust their respective amplitude variations (e.g., beatings) in the synthetic signals. Adjusting the residues cannot solve this problem because the residues are not allowed to vary in time.

Careful selection of the original signals. The best way of avoiding the problem mentioned above remains to choose original signals whose sinusoidal frequencies do not overlap! This can be tricky for the piano, because all tones are quasi-harmonic and the tempered scale involves nearly-rational multiplicative factors. The multiplicative ratios corresponding to the most critical intervals are given below, with their rational approximation and the corresponding error:

We see that the most critical intervals are (in order of 'criticalness'): the fourth, the fifth, the third and the sixth. For example, in an interval of a fifth ($\approx \frac{3}{2}$), the second harmonic of the upper tone and the third harmonic of the lower tone are only 0.11% apart in frequency.

Interval	Exact Ratio	Approximation	Error
3^{rd}	1.25992	5/4	0.78%
4^{th}	1.33483	4/3	0.11%
5^{th}	1.49830	3/2	0.11%
6^{th}	1.68179	5/3	0.9%

Table 1: Intervals, corresponding ratios, rational approximations and errors.

The example above corresponded to an interval of a fourth. The component with the biased amplitude was the upper tone's third harmonic ($2111Hz$) which coincided with the fourth harmonic of the lower tone ($2107Hz$).
If we want to avoid the three most critical intervals (the fourth, the fifth and the third), then we can only use four notes per octave (e.g., C, $D^\sharp$, $F^\sharp$, A). We can also remark that the third interval has a less accurate rational approximation than the fourth and the fifth intervals: avoiding only those two intervals makes it possible to simultaneously use six notes per octave (arranged on a tone scale: e.g., C, D, E, $F^\sharp$, $G^\sharp$, $A^\sharp$ or $C^\sharp$, $D^\sharp$, F, G, A, B).
Indeed, when such series are used, original and synthetic tones are nearly indistinguishable. Only the background noise (tape hiss) disappears: the tape hiss present in all the original recordings can be considered non-correlated. A single source couldn't possibly generate independent noises, when fed through different resonant filters.

V. Conclusion and prospects

A. Application to the piano

From a practical point of view, the application of our models to piano tones is very promising. Below are some of the main results and topics that require further investigations.
A low-cost technique for the synthesis of very realistic piano tones. Modern synthesizers mostly rely on sampling techniques for the synthesis of realistic piano tones. Sampling techniques require large amounts of memory to store the recordings of successive piano notes (even when only one note out of four or five is sampled) at different velocities. To minimize the memory requirements, only a short segment of the quasi-stationary part of the tones is usually recorded, and looped during play-back (with an exponential weighting). This gives birth to periodic audible artifacts due to looping cross-fading.
Our model simplifies significantly the synthesis of high-quality piano tones by dividing the memory requirements by at least 6. The resonant filters are simple IIR filters which can easily be implemented on standard DSP chips. For example, the biquad implementation on a standard Motorola DSP56000 requires between 6 and 10 cycles per sample per second-order cell: over 25 sections can be implemented simultaneously at a sampling rate of 48kHz on a 25MHz DSP56000, or over 50 sections at a sampling rate of 32kHz on a 33MHz DSP56000.
A more natural onset. One problem with current piano synthesizers is the way the multiple strokes are handled: in sampled pianos, when the key is struck again, the sound is simply played again (with a smooth transition). This scheme is quite different from what happens in a real piano and in particular, there is no possibility of 'sound build-up' which occurs in real pianos when the same key is struck repetitively.
Our model on the contrary supports this kind of effect: the excitation can be fed again into the resonant filter and generate a new resonance that is simply superimposed to the preceding

one. The result is a very natural-sounding repetition of the note: each stroke generates a new note that does not exactly sound like the preceding ones because the resonant filter has different initial conditions.

Improvements: sympathetic resonances. Another effect that is currently not properly rendered by synthetic pianos is the sympathetic-resonances effect. When a string is vibrating, its vibration is transmitted to the soundboard which in turns excites the whole set of strings. If some other strings are not damped, and if their natural resonating frequencies can be excited by the originally vibrating strings, they start vibrating at their own frequency in an audible effect know as sympathetic resonance. This is always the case for the highest notes (the upper octave) of the piano, and is true for all notes when the sustain pedal is held down.

This effect can be easily incorporated in our model by feeding a small part of the active filter's output into resonant filters corresponding to sympathetic strings, very much like what actually happens in a real piano. This cross-feeding could also include a transfer function designed to model the sound path from the vibrating string to the sympathetic strings (bridge, soundboard) [29, 30].

Improvements: shortening the excitation. To further minimize the memory requirements in our model, two things can be done:

1. We can save only a small part of the common excitation signal (say four or five seconds), and damp its end towards zero. When the excitation drops to zero, the filter continues resonating on its own, with a damping that corresponds to the original sound. In this case, the beatings can be lost if the resonant filter does not include them (see part C).

2. we can save only a small part of the common excitation signal, and loop it during the synthesis, with appropriate cross-fading and weighting (as is done in sampled pianos). Although looping artifacts are quite audible in sampled pianos (resulting from cross-fading), syntheses involving looping of the excitation signal are *free of artifacts*. The periodic cross-fading is audible in the excitation signal, but generates no audible distortion in the output of the resonant filter.

We have tried both methods with similar results in terms of tone quality. The first, simplest method can be used for higher notes which decay rapidly: in the original tones, the beatings disappear after only one or two seconds and can therefore be left in the excitation signal. The second method can prove useful for the lower notes (below C2) which last much longer. Very low notes can be perfectely synthesized by using a truncated excitation and resonant filters that include beatings.

Improvements: Simplifying the resonant filter. Although the highest notes of the piano have only a few harmonics, the lowest notes are made of a large number of sinusoidal components: the note $A1$ (55 Hz) contains over 60 visible harmonics! As a result, the resonant filters corresponding to low frequencies are very complex and time-consuming when implemented as described in part C (sum of second-order sections). An efficient way of reducing this complexity consists in replacing the original resonant filter by a modified comb-filter [31, 32]. The comb-filter's feedback loop includes a rational-transfer function filter of low-order as shown in Fig. (17) By use of identification methods [30], it is possible to design a modified comb-filter that closely matches the quasi-harmonic structure of the original resonant filter and requires much fewer

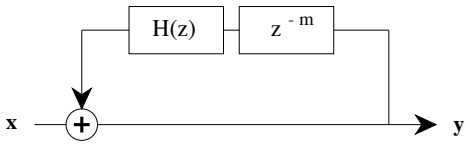

Figure 17: A modified comb-filter alternative to high-order resonant filters.

calculations than the corresponding sum of second-order sections.

This more efficient filter can be used for the lowest notes of the piano, and a sum of second-order sections implemented for the medium and high notes.

B. Theoretical results

The source/filter models have the feature of separating the excitation signal from the resonance. This makes it possible to work on the excitation signal whose structure is more difficult to characterize than that of the resonance. We have seen that obtaining an excitation signal from a real sound is sometimes a delicate task that can suffer from ill-conditioning. However, the examples presented above demonstrated the good stability of our calculation methods. In particular, it was interesting to observe that in the case of a piano, the specific excitations corresponding to adjacent notes appear very similar in many ways (Figs. (8) and (12)). This result tends to validate the analogy 'physical-exciter $\leftrightarrow$ excitation-signal', 'instrument-resonator $\leftrightarrow$ resonant-filter' and justifies the application of the single-excitation/multiple-filter model.

These models have also been applied to the guitar with very good results. The excitation signals exhibit a great uniformity across several notes, and the signal-excitation/multiple-filter model yields very accurate syntheses.

In summary, our experiments demonstrate that:

1. Percussive sounds can be modeled as the output of resonant systems excited by an excitation signal.

2. The excitations corresponding to adjacent notes remain very similar.

3. As a consequence, it is possible to find an excitation that is common to up to 6 or 7 notes belonging to the same octave (depending on the instrument).

These results in themselves are very promising and call for further research. Our models need to be applied to other musical instruments (e.g., vibraphone, bells) to test their generality. In addition, we need to investigate a number of questions:

Variation of excitation with velocity: it is well known that the energy of percussive instrument sounds increases more in high frequencies when the velocity of the physical exciter increases (see [27] for the piano). Since the resonant structure remains the same whatever the velocity, this change of frequency content should be observed in the excitation signal.

Variation of excitation from one octave to the other: we have seen for the piano that the excitation remains about the same for notes belonging to the same octave. The next step would be to determine how the excitation signal changes from one octave to the next one, for different kinds of music instruments.

Link between calculated excitation and 'physical' excitation: By using measurements of the force at the hammer at the moment of impact, it is possible to gain access to what

could be called a physical 'excitation' signal. An important question remains as to how this physical excitation signal compares to our excitation signal.

Model of the excitation: is it possible to model the common excitation itself? More precisely, we would like to be able to model the excitation signal, its dependance on the velocity and its octave-variations. This, we believe, is the most challenging problem.

The authors would like to thank J.M. Jot for his helpful discussions and the reviewers for their insightful comments.

Appendix

A. Appendix A

In this appendix, we show that the filter $H(z)$ defined by

$$H(z) = \sum_{i=1}^{p} \frac{r_i}{1 - z_i z^{-1}}$$

with r_i *real and positive* has no zero outside the disk $\{|z| < \max_i |z_i|\}$. In other words, the zeros of filter $H(z)$ are of modulus smaller than the maximum of the pole moduli.

If we define

$$h_i(z) = \frac{r_i}{1 - z_i z^{-1}}$$

then

$$Real(h_i(z)) = \frac{r_i \ (1 - Real(z_i z^{-1}))}{|1 - z_i z^{-1}|^2}$$

which is strictly positive if $|z| > |z_i|$. This proves that

$$Real(H(z)) = \sum_{i=1}^{p} Real(h_i(z))$$

is also strictly positive if $|z| > \max |z_i|$. Therefore $H(z)$ cannot have zeros outside the disk of radius $\max |z_i|$.

An interesting corollary is obtained by setting $r_i = 1$, in which case

$$H(z) = \frac{z \ A'(z)}{A(z)}$$

with

$$A(z) = \prod_{i=1}^{p} (z - z_i)$$

The preceding result shows that if polynomial $A(z)$ has all its roots inside a disk, then the roots of its derivative $A'(z)$ all lie inside the same disk.

B. Appendix B

In this appendix, we derive the expression of the common excitation signal.

The problem is the following:

Given p original signals x_n^i $(0 \leq i < p)$ and p resonant filters H_i with impulse responses h_n^i we search for an excitation signal e_n which minimizes the cumulative error $\mathcal{E}$ defined as:

$$\mathcal{E} = \sum_{i=0}^{p-1} \left(\sum_{k=0}^{N-1} (x_k^i - \hat{x}_k^i)^2 \right) \qquad \text{with}$$

$$\hat{x}_k^i = h_n^i \star e_n$$

First, Parseval's theorem allows us to write the preceding equation in the frequency domain:

$$\mathcal{E} = \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\sum_{i=0}^{p-1} | \ X_i(f) - H_i(f) E(f) \ |^2 \right) df$$

in which $X_i(f)$ is the Fourier transform of x_n^i, $H_i(f)$ is the Fourier transform the impulse response h_n^i, and $E(f)$ is the Fourier transform of the excitation signal.

Now minimizing the integral defining $\mathcal{E}$ amounts to minimizing the integrant with respect to $E(f)$ for every value of f:

$$\min_{E(f)} \sum_{i=0}^{p-1} | \ X_i(f) - H_i(f) E(f) \ |^2$$

This is now a standard complex quadratic minimization whose solution is given by

$$\sum_{i=0}^{p-1} (H_i^\star(f)(X_i(f) - H_i(f)E(f))) = 0$$

or

$$E(f) = \frac{\sum_{i=0}^{p-1} H_i^\star(f) X_i(f)}{\sum_{i=0}^{p-1} H_i^\star(f) H_i(f)}$$

The time domain excitation signal is then obtained by an inverse Fourier transform.

References

[1] L. Meirovitch, *Elements of Vibration Analysis*, McGraw-Hill, New York, 1986.

[2] P.M. Morse and K. U. Ingard, *Theoretical Acoustics*, McGraw-Hill, New York, 1968.

[3] J. Laroche, *Etude d'un Système d'Analyse et de Synthèse Utilisant la Méthode de Prony. Application aux Instruments de Musique de Type Percussif*, PhD thesis, ENST, Oct 1989.

[4] J. A. Moorer, "Signal processing aspect of computer music", *Proc. of the IEEE*, vol. 65, Aug 1977.

[5] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice Hall, Englewood Cliffs, New Jersey, 1975.

[6] A. Chaigne, "On the use of finite differences for musical synthesis. application to plucked stringed instruments", *J. d'acoustique*, vol. 5, pp. 181–211, Apr 1992.

[7] J. D. Markel and A. M. Gray, *Linear prediction of speech*, Springer-Verlag, Berlin, 1976.

[8] G. Poirot, X. Rodet, and P. Depalle, "Diphone sound synthesis based on spectral envelopes and harmonic/noise excitation functions", *Proc. of International Computer Music Conference, Köln*, Sep 1988.

[9] P. Depalle, *Analyse, Modélisation et synthèse des sons basées sur le modèle source-filtre*, PhD thesis, Université du Maine, Le Mans, France, Dec 1991.

[10] Y. Potard, P. F. Baisnée, and J. B. Barrière, "Experimenting with models of resonance produced by a new technique for the analysis of impulsive sounds", *Proc. of International Computer Music Conference*, 1986.

[11] J. B. Barrière, A. Freed, P. F. Baisnée, and M. D. Baudot, "A digital signal multiprocessor and its musical application", *Proceedings of 1989 International Computer Music Conference, Columbus, Ohio*, 1989.

[12] J. Laroche, "The use of high resolution methods for the analysis of musical signals", *Journal of the Acoustical Society of America*, Submitted for publication 1992.

[13] R. O. Smith, "Multiple emitter location and signal parameter estimation", *Proc. RADC Spectrum Estimation Workshop (Rome, NY)*, pp. 243–258, 1979.

[14] J. Laroche, "A new analysis/synthesis system of musical signals using prony's method. Application to heavily damped percussive sounds", *Proc. IEEE ICASSP-89, Glasgow*, pp. 2053–2056, May 1989.

[15] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation", *IEEE Trans. Acoust., Speech, Signal Processing.*, vol. ASSP-34, pp. 744–754, Aug 1986.

[16] M.R. Schroeder, "New method for mesuring reverberation time", *Journal of the Acoustical Society of America*, vol. 37, pp. 232–235, 1965.

[17] J. M. Jot, "An analysis/synthesis approach to real-time artificial reverberation", *Proc. IEEE ICASSP-92, San Francisco*, Mar 1992.

[18] J. Makhoul, "Linear prediction: A tutorial review", *Proc. of the IEEE*, vol. 63, pp. 1380–1418, Nov 1975.

[19] G. Demoment, "Image reconstruction and restoration: Overview of common estimation structures and problems", *IEEE Trans. Acoust., Speech, Signal Processing.*, vol. ASSP-37, pp. 2024–2036, Dec 1989.

[20] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech", *Proc. of the IEEE*, vol. 67, Dec 1979.

[21] L. L. Scharf, *Statistical Signal Processing*, Addison-Wesley, New York, 1991.

[22] A. M. Thompson, J. C. Brown, J. W. Kay, and D. M. Titterington, "A study of methods of choosing the smoothing parameter in image restoration by regularization", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 326–339, Apr 1991.

[23] M. Z. Nash, *Generalized Inverses and Applications*, Academic Press, New York, 1976.

[24] M. P. Ekstrom, "A spectral characterization of the ill-conditioning in numerical deconvolution", *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 344–348, Aug 1973.

[25] D. L. Wessel, "Low dimensional control of musical timbre", Technical report, IRCAM, Paris, dec 1978.

[26] D. L. Wessel, "Psychoacoustics and music: A report from michigan state university", *PAGE: Bulletin of the Computer Arts Society.*, vol. 30, 1973.

[27] A. Askenfelt and E. Jansson, "From touch to string vibrations", in A. Askenfelt, editor, *The Acoustics of the Piano*, pp. 39–57. Royal Swedish Academy of Music, Stockholm, 1990.

[28] G. Weinreich, "Coupled piano strings", *Journal of the Acoustical Society of America*, vol. 62, Dec 1977.

[29] K. Wogram, "Acoustical research on pianos: Vibrational characteristics of the soundboard", *Das Musicinstrument*, pp. 694–702,776–782,872–880, 1980.

[30] J. O. Smith, *Techniques for Digital Filter Design and System Identification with Application to the Violin*, PhD thesis, Stanford University, Stanford, CA, Jun 1983.

[31] D. A. Jaffe and J. O. Smith, "Extensions of the karplus-strong plucked-string algorithm", *Computer Music Journal*, vol. 7, pp. 56–69, Summer 1983.

[32] J. Laroche and J.M. Jot, "Analysis/synthesis of quasi-harmonic sounds by use of the karplus-strong algorithm", *Proceedings of the 1992 SFA conference, Arcachon*, Apr 1992.