Automatic Piano Transcription Using Frequency and Time-Domain Information

Juan P. Bello, *Member, IEEE*, Laurent Daudet, *Member, IEEE*, and Mark B. Sandler, *Senior Member, IEEE*

***Abstract*—The aim of this paper is to propose solutions to some problems that arise in automatic polyphonic transcription of recorded piano music. First, we propose a method that groups spectral information in the frequency–domain and uses a rule-based framework to deal with the known problems of polyphony and harmonicity. Then, we present a novel method for multipitch-estimation that uses both frequency and time–domain information. It assumes signal segments to be the linearly weighted sum of waveforms in a database of individual piano notes. We propose a solution to the problem of generating those waveforms, by using the frequency–domain approach. We show that accurate time–domain transcription can be achieved given an adequate estimation of the database. This suggests an alternative to common frequency–domain approaches that does not require any prior training on a separate database of isolated notes.**

***Index Terms*—Audio, F0 estimation, music, multiple pitch estimation.**

I. Introduction

WE CAN define music transcription as the process of converting a musical recording or performance into a musical score, or equivalent representation. In the traditional sense, transcribing a piece of music implies a number of high-level tasks such as: estimating the pitch and timing of individual notes; estimating the tempo, meter and key of each section of the piece; identifying and labeling ornamentations; recognizing the instruments being played; and segregating "voices" according to the instrument that played them and to their function, i.e., melody, accompaniment, etc.

These tasks, already complicated for highly trained individuals such as musicians and musicologists, have proven extremely difficult for computers. Relatively successful methods have been proposed for monophonic signals, i.e., when only one note is present at a time. However, success has been more evasive for the analysis of polyphonic signals, i.e., presenting a multiplicity of notes and, possibly, instruments at a time.

At its simplest, the issue of transcribing music, except for information related to timbre and instrumentation, can be reduced to knowing the fundamental frequency (f_0), start time, duration, and loudness of individual musical notes in the recording. We can, therefore, assume that note event information is enough to describe a musical signal such that this encoding can be used for a range of high-level applications including: retrieval of musically similar recordings from large audio databases, coding of audio information for fast audio transmission through data channels (e.g., MPEG-4 coders), real-time high-level interaction between musicians and computers, analysis of recordings of the same piece by different performers, etc.

This paper aims to propose novel ways of automatically extracting note events from simple polyphonic audio files, i.e., real recordings including an undetermined number of notes at a time played by a single instrument. Specifically, we have chosen the piano as our single instrument, because this is one of the instruments where the problems due to polyphony are the most challenging. Also, there exists a large corpus of solo piano music that can be used for testing and evaluation. Finally, we can exploit the characteristics of the piano sound to our benefit, as will be seen in later sections.

Although this is not a transcription system in the strict musical sense, we will indistinctly refer to our results as transcriptions and to the research area as automatic music transcription.

A. Background

Almost invariably, since the early works by Moorer [1] and Piszcalski and Galler [2], polyphonic music transcription systems rely on the analysis of information in the frequency domain: Klapuri [3] uses the iterative calculation of predominant f_0s in separate frequency bands; Martin [4] uses blackboard systems; and Kashino *et al.* [5] use Bayesian probability networks for the grouping of supportive frequency–domain evidence. Raphael proposes the use of a hidden Markov model and spectral feature vectors to describe chord sequences in piano music signals [6]. Carreras *et al.* [7] use neural networks for spectral-based harmonic decompositions of signals, while Marolt [8] uses networks of adaptive oscillators to track partials over time. A recent example, Ortiz *et al.* [9], uses a physical model of the piano to generate spectral patterns that can be compared to the incoming spectral data. For an extensive review of polyphonic music transcription systems, see [10].

Analyzing spectral data is justified as in time-frequency representations, periodicities in time are represented as energy maxima (i.e., peaks) in the frequency–domain. This suggests that principled grouping of these energy maxima generates patterns or structures that may be related to notes in a music signal. Notably, the presence of a note is specifically associated with the presence of a comb-pattern in the frequency–domain with lobes approximately at the positions of the multiples of the fundamental frequency of the analyzed tone.

However, relying on the analysis of the frequency–domain data has some disadvantages [10], [11]. The resolution limitations of most time–frequency representations can badly affect

Manuscript received March 2, 2005; revised October 8, 2005. This work was supported in part by the European Commission through the SIMAC Project IST-FP6-507142. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerald Schuller.

J. P. Bello and M. B. Sandler are with the Centre for Digital Music Department of Electronic Engineering, Queen Mary, University of London, London E1 4NS, U.K. (e-mail: juan.bello-correa@elec.qmul.ac.uk; mark.sandler@elec.qmul.ac.uk).

L. Daudet is with the Laboratoire d'Acoustique Musicale, Université Pierre et Marie Curie (Paris 6), 75015 Paris, France (e-mail: daudet@lam.jussieu.fr).

Digital Object Identifier 10.1109/TASL.2006.872609

the frequency localization of spectral components. This, in turn, is emphasized by the polyphony of the signal: when more than one note is present, peaks related to different f_0s can lie in the same frequency bin and, therefore, are not uniquely identifiable. The problem becomes worse when the polyphony increases, often increasing the number of estimation errors. Overlapping between the different harmonic (or nearly harmonic) series is also produced by the presence of harmonic relations between sounds (i.e., harmonicity). This phenomena sharply diminishes the chances of resolving the mixture of notes in the signal.

B. Organization of This Paper

In the following sections, we present a method that uses the commonly used approach of grouping frequency–domain data for polyphonic music transcription (Section II) with some novel adaptations by means of an heuristic set of rules. Then, we propose that we can avoid the paradigm of analysis only in the frequency–domain, by using a time–domain linear additive approach (Section III) which is well suited to work with piano signals. We integrate frame-by-frame results into note events (Section IV) and perform a comparative study of the two methods on a database of polyphonic piano recordings (Section V), finally presenting the conclusion and further considerations (Section VI).

II. Frequency–Domain Transcription

The goal of our frequency–domain method is to find groups of spectral peaks that characterize the notes in a recorded mixture. To this end we will calculate the signal's spectrum by means of the short-time Fourier transform (STFT) and select strong peaks in the frequency–domain, group selected peaks according to expected harmonic comb patterns, and use a number of heuristic rules to select the comb patterns that best describe the fundamental frequencies (f_0) that exist in a given signal segment.

A. Spectral Peak-Picking

Let us consider the STFT of the signal $s(n)$

$$S(k) = \sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1} s(nh+m)w(m)\mathrm{e}^{-\frac{2j\pi mk}{N}} \tag{1}$$

where $w(m)$ is an N-point window, and h is the hop size, or time shift, between adjacent windows. The frequency resolution is $\Delta f = f_s/N$, where f_s is the sampling frequency.

The magnitude of $S(k)$ includes a number of irrelevant peaks that may mislead the estimation procedure. In [12], a psychoacoustic masking model is used to eliminate weaker peaks. In this paper, we smooth the spectrum using a zero-shift infinite impulse response (IIR) digital filter [13]. Peak-picking is performed on the filtered spectrum using a simple local maximum algorithm, such that detected peaks correspond to the stronger peaks in the original spectrum. Peaks are matched to the closest local maximum of the unfiltered spectrum to compensate for the loss of resolution brought about by the filtering. The resulting modified spectrum $S_p(k)$ of detected peaks can be defined as

$$S_p(k)\begin{cases} |S(k)| & \forall k \in \left[0, \left(\frac{N}{2}\right)-1\right], \text{if peaks are detected} \\ 0, & \text{elsewhere.} \end{cases}$$

B. Grouping Peaks

Pitched sounds are expected to produce frequency–domain comb patterns with lobes approximately at mf_0, $m = 1, \ldots, M$, where f_0 is the fundamental frequency of the sound, and $M = \min\{M_{\max}, fs/2f_0\}$ is the number of lobes in the pattern. These patterns are known as harmonic combs. Here, because our approach is energy-based, and since most of the energy is concentrated on the first few partials, we can neglect, as a first-order approximation, the well-known (and perceptually important) effect of piano strings inharmonicity.

To identify these combs, we need first to generate a list of possible fundamental frequencies. A possible approach [14] is to select the P spectral peaks with greater magnitude, and then, for each selected peak p, to generate Z note hypotheses with fundamental frequencies defined by

$$f_0(p,z) = \frac{f_i(k_p)}{z}, z \in [1, Z] \tag{2}$$

such that

$$k_p \equiv \text{bin of the } p\text{th maximum of } S_p(k), p \in [1, P] \tag{3}$$

and $f_i(k), k \in [0, N/2-1]$, is the bin instantaneous frequency calculated using the phase-vocoder technique [15], [16], i.e., the bin's unwrapped phase difference divided by the STFT's hop size. This improves the precision of the frequency estimation. We can use $f_0(p,z)$ to generate a $P \times Z$ matrix of possible harmonic combs in the spectrum $S_p(k)$.

Let us define k_m, a subset of $k \in [0, N/2-1]$, such that $\forall k \in k_m$ the instantaneous frequencies $f_i(k) \in [f_m 2^{-1/24}, f_m 2^{1/24}]$ are within a quarter tone distance from $f_m = mf_0(p,z)$, the expected frequency of the mth comb's lobe. The hypothesis associated with this comb, $H_{p,z}(m)$ can be defined as

$$H_{p,z}(m) = \max\{S_p(k)\}, k \in k_m, m \in [1, M]. \tag{4}$$

The array $H_{p,z}(m)$ represents the closest approximation between the ideal combs defined by $mf_0(p,z)$ and the spectral data $S_p(k)$. An example can be seen in Fig. 1, where the harmonic grids for $p = 1$, $z = 1$,3 and $M_{\max} = 7$ are selected. In our implementation $M_{\max}$, P, and Z are static predefined values, experimentally set to 12, 10, and 5, respectively.

C. Selecting Comb Patterns

We need to evaluate the $P \times Z$ array H of observed comb patterns, and select those hypotheses that best describe the f_0s in a given signal segment. Selection is performed by a process of elimination of weak hypotheses, i.e., those that are related to a few low-magnitude peaks or that can be explained by the combination of other hypotheses. In the following, we provide a list of heuristic rules applied sequentially to all hypotheses. All

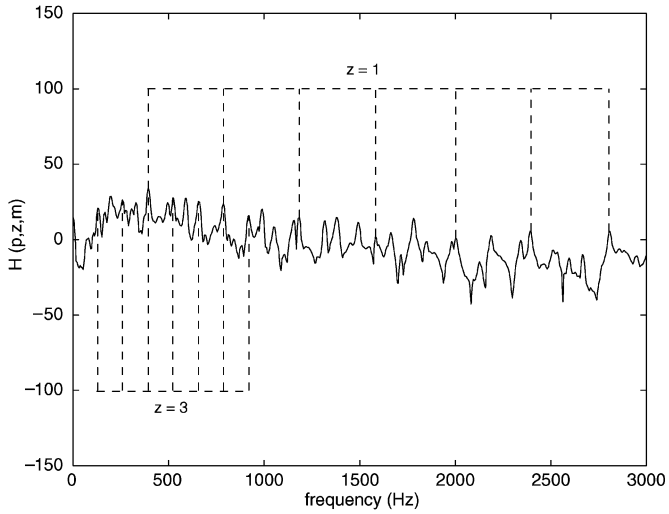

Fig. 1. Harmonic combs for two different roots ($z = 1$ and $z = 3$) that include the maximum peak of $S(k)$.

used thresholds are static and set experimentally. For a comb pattern to be selected, all tests need to be satisfied.

1) Minimum Support: When constructing H, peaks are searched for in the vicinity of the expected position of a partial. If no spectral peaks are found within that region, it is said that there is no support from that partial. For a hypothesis to be selected, it requires a minimum number of detected partials (relative to its f_0) which is at least equal to $M/2$.

2) Minimum Energy: The sum of the energy of all supporting partials (i.e., total energy of the comb) must be above a minimum energy threshold equal to 10 dB.

3) Detection of Subharmonics: Consider an octave-related pair of hypotheses, H_{high} and H_{low}, such that their fundamental frequency ratio $f_{\text{high}}/f_{\text{low}}$ is approximately an even number. Thus, the even partials of f_{low} overlap all partials of f_{high}. If any of the followings observations is true, then H_{low} is considered a subharmonic benefiting from the energy of H_{high} and therefore eliminated: First, that the energy produced by the even partials of the subharmonic is significantly higher (three times higher in our implementation) than that produced by its odd partials

$$\sum_{m\ \text{odd}}^{M} [H_{\text{low}}(m)]^2 \ll \sum_{m\text{even}}^{M} [H_{\text{low}}(m)]^2 \tag{5}$$

where M is the total number of partials. Second, that the energy of the initial $J(= M/4)$ partials of the lower note is significantly smaller (eight times smaller in our implementation) than its total energy

$$\sum_{m}^{J} [H_{\text{low}}(m)]^2 \ll \sum_{m}^{M} [H_{\text{low}}(m)]^2 . \tag{6}$$

4) Detection of Overtones: For the same octave-related pair, let us analyze the distribution of the energy through all the comb lobes. If the total amount of the energy of H_{high} is concentrated on its first J partials

$$\sum_{m}^{J} [H_{\text{high}}(m)]^2 \approx \sum_{m}^{M} [H_{\text{high}}(m)]^2 \tag{7}$$

then H_{high} is considered an overtone and thus eliminated.

5) Harmonic Overlapping: Two notes in harmonic relation share partial components according to the relationship $af_1 = bf_2$, where a and b are integer values. If considering harmonic intervals other than octaves, it is possible that a hypothesis H_x exists only as a result of the presence of other harmonically related hypotheses $(H_1, \ldots, H_N)$ in the signal segment. For example, if a chord $E_4(329.628\,\text{Hz}) + G_4(391.995\,\text{Hz})$ is played, peaks will be generated approximately at the position of every third and fifth partial of C_4 (261.626 Hz). That is five of the first ten partials of $C4$, therefore generating strong support from existing peaks. In this case, we can define

$$h = \sum_{m}^{M} m \left(\sum_{i=1}^{N} a_i f_x \right) = \sum_{m}^{M} m \left(\sum_{i=1}^{N} b_i f_i \right) \tag{8}$$

a subgroup of partials of the comb H_x that overlaps with partials from the harmonically related hypotheses $H_1, \ldots, H_N$. If H_x is only the result of these harmonic relationships, then its total energy should be almost equal to the energy of the subgroup h (an inverse criteria can be defined for $m \notin h$)

$$\sum_{m}^{M} [H_x(m)]^2 \approx \sum_{m\in h}^{M} [H_x(m)]^2 . \tag{9}$$

If true, H_x is considered a hypothesis generated by harmonicity and thus eliminated.

6) Competitive Energy: Let us define $H_{\max}$ as the hypothesis with the highest energy. All remaining hypothesis H_i must comply with

$$\gamma \sum_{m}^{M} [H_i(m)]^2 \geq \sum_{m}^{M} [H_{\max}(m)]^2 \tag{10}$$

where $\gamma(= 0.1)$ is a predefined value. If this condition is false, then H_i is considered insignificant and thus eliminated.

As will be seen in Section V, this method gives relatively good results (about 69% of notes are correctly detected, for 16% of false positives). However, at high levels of polyphony, i.e., chords of more than four notes, it suffers from the usual limitations of the analysis in the frequency domain: too many partials are intertwined together, and typical errors such as octave errors are prone to appear.

This frequency–domain method could be improved in many ways. For instance, we could take into account a standard inharmonicity law [17] to estimate more precisely the frequency of partials: $f_n = nf_0\sqrt{(1 + B(n-1)^2)}$, where B is a constant that depends on the physical characteristics of the string. However, values for B are strongly note- and piano-dependent, so this hyper-parameter would have to be learned from the signal.

This physically motivated improvement could help us resolve some chord ambiguities, but at a cost of a large increase in the complexity of the model. Instead, we have chosen to enhance our signal model by combining this frequency–domain approach with a new time–domain method, which will be the focus of the next sections.

III. Time–Domain Transcription

We have seen that almost all existing methods for polyphonic transcription operate only on frequency–domain data. In this section, we propose an alternative approach that avoids, at least partially, the usual paradigm of analysis in the frequency domain. The system uses a hybrid method,[1] where the classical approach is improved by a time–domain recognition process. This enables a refinement of our results by taking into account the information contained in phase relationships, that are lost when only the magnitude spectra of sounds are analyzed. The method is applied to piano music, where, as a first-order approximation, the phase relationships between partials in a given piano note can be assumed reproducible. This assumption is very specific to instruments where the player has little or no control on the excitation (apart from velocity), and where the exact location of the instrument does not vary in time.

A. Linear Additive Approach

Let us return to $s(n), n = 1, \ldots, N$, a segment of our signal. Let x_i, $i = 1, \ldots, L$, be the time–domain normalized waveform of one of the individual notes of a single instrument (for our implementation to piano music $L = 88$). Let us assume (as an initial approximation) that each x_i is independent of its loudness, that is, the waveform remains the same regardless of the strength at which the corresponding note has been played, except for a global scaling of the signal's amplitude.

Let us also assume that, in a mixture, a waveform produced by a given single note is independent of the presence of other waveforms. In other words, we neglect the interactions due to mechanical coupling phenomena that may arise when two or more keys are pressed at the same time.

Furthermore, let us demonstrate that the individual waveforms x_i form a family of linearly independent vectors. This means that it is not possible to obtain a note by a linear combination of other notes, as can be shown by contradiction: Let us assume linear dependency between the individual notes x_i, such that $\sum_i a_i x_i = 0$, for a_i not all equal to zero. Let i_0 be the smallest index, if any, such that $a_{i_0} \neq 0$. The signal $x_{i_0}(n)$ contains the fundamental frequency corresponding to the i_0th note (the fundamental is always present in piano sounds), which is not present in any of the other x_i, $i > i_0$; therefore, it is necessary that $a_{i_0} = 0$, hence, the contradiction.

Let $\mathcal{D} = \{x_i\}_{i=1,\ldots,L}$ be the database containing the L waveforms of the individual notes. In this context, the resulting waveform $s(n)$ of a chord of synchronous notes can be simply defined as a weighted linear sum of the individual notes x_i

$$s(n) = \sum_{i=1}^{L} \alpha_i x_i(n) \tag{11}$$

[1]Despite this, and to simplify the comparison between the two approaches, we will refer to this method simply as the "time–domain" approach.

where α_i is the mixing coefficient for the ith note, such that

$$\alpha_i \begin{cases} > 0, & \text{if the } i \text{ th note is played in } s(n) \\ = 0, & \text{otherwise.} \end{cases}$$

where α_i increasingly maps the loudness, or velocity, of the corresponding note.

Following this definition, the frame-by-frame f_0-estimation problem, can be restated as the calculation of the values of the mixing vector $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1,\ldots,L}$, given the segment $s(n)$ and the database $\mathcal{D}$. This operation returns information about what notes have been played in the segment and with what loudness.

In finite dimensions, a simple algebraic solution can be found to this inverse problem. Let us define D, the representation of $\mathcal{D}$ on $\mathbb{R}^N$, a $L \times N$ matrix whose rows are the N-length individual vector notes x_i

$$D = \begin{bmatrix} x_1(1) & x_1(2) & \cdots & x_1(N) \\ x_2(1) & x_2(2) & \cdots & x_2(N) \\ \vdots & \vdots & \vdots & \vdots \\ x_L(1) & x_L(2) & \cdots & x_L(N) \end{bmatrix}.$$

Equation (11) can now be expressed in terms of D as

$$s = D^T \boldsymbol{\alpha} \tag{12}$$

and, therefore

$$Ds = DD^T \boldsymbol{\alpha}. \tag{13}$$

Because of the linear independence of the rows of D, the $L \times L$ matrix DD^T is not singular, thus invertible. Hence, (13) is equivalent to

$$\boldsymbol{\alpha} = (DD^T)^{-1} Ds. \tag{14}$$

Therefore, the mixing vector $\boldsymbol{\alpha}$ can be reconstructed by a simple matrix product of the fixed matrix $(DD^T)^{-1}D$ and the segment s. The rows of $(DD^T)^{-1}D$ form the dual basis $\mathcal{D}^\star$ of the basis $\mathcal{D}$; $\boldsymbol{\alpha}$, which represents the orthogonal projection of s on the subspace $\mathcal{D}$, is obtained by scalar products with elements of $\mathcal{D}^\star$. It is important to note that this is a simultaneous estimation of all notes, as opposed to the standard recursive processes of the majority of frequency–domain techniques. We will discuss the details of the estimation of D in Sections III-C and III-D.

B. Phase Alignment

Phase-alignment between simultaneous notes is only possible under specific conditions, i.e., audio files generated with synthesized sounds using perfectly quantized playing. However, this is not the case when working with real recordings. Individual notes within a chord are never perfectly synchronous. Thus, the above-mentioned approach is over-simplified (as all scalar products assume alignment), and the results obtained using this method are not accurate.

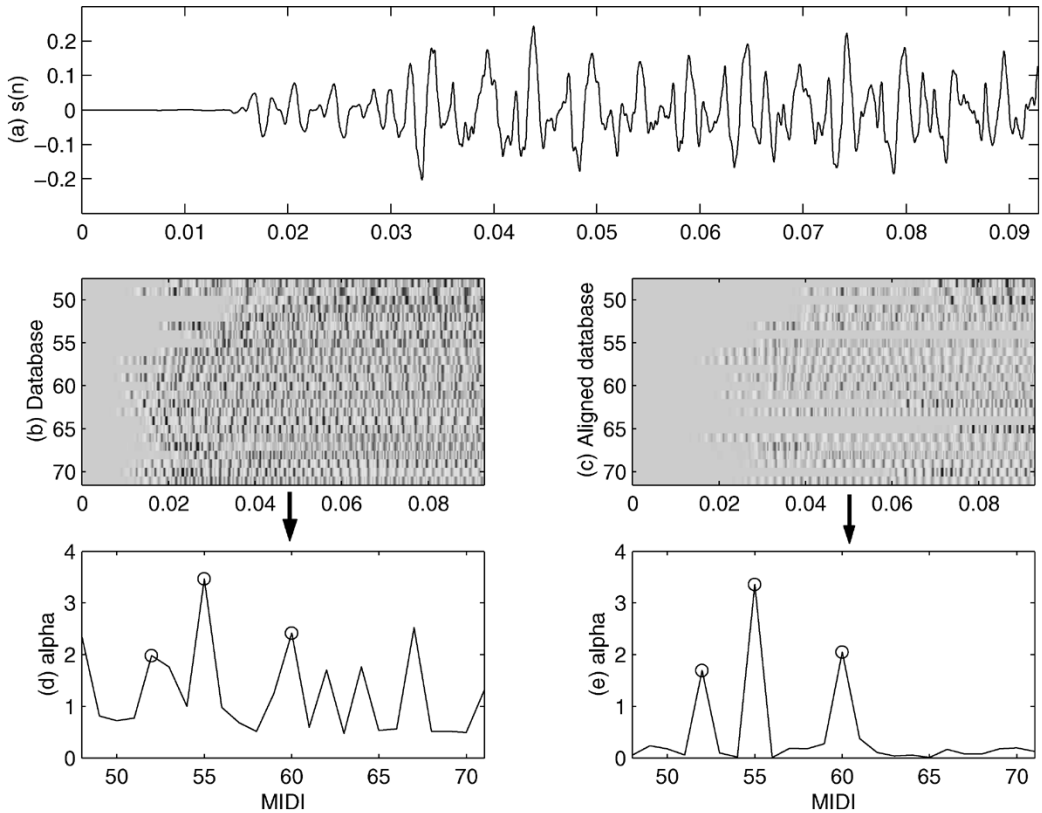

Fig. 2. (a) Estimation of the weighting coefficients $\boldsymbol{\alpha}$ on a segment $s(n)$, (b) using the nonaligned database D, and (c) the aligned database $\tilde{D}$. Estimation of $\boldsymbol{\alpha}$ is more accurate (e) with alignment than (d) without (circles indicate the actual notes).

Let us define τ_t, the shift-by-t-samples operator, such that

$$\tau_t x_i(n) = x_i(n-t). \tag{15}$$

A possible solution for our phase-alignment problem is to consider all shifted versions of vectors x_i up to a delay T, or $\tau_t x_i$, $t = 1, \ldots, T$. This means that we will work in subspaces of higher dimension (such as $L \times T$).

Needless to say, this is a very computationally intensive solution. Moreover, when $L \times T$ gets larger than N (as will happen for large values of T) the family of vectors becomes overcomplete; hence, it is no longer linearly independent. This implies the noninvertibility of matrix DD^T, thus precluding the calculation of vector $\boldsymbol{\alpha}$ as stated in (14). An alternative approach is needed for phase alignment.

On a frame-by-frame basis, let us suggest that for the ith note in the database, we only need to use the $\tau_{t_i} x_i$, such that t_i compensates for the phase misalignment between the sound $s(n)$ and the note x_i. This is equivalent to generalizing (11) as

$$s(n) = \sum_{i=1}^{L} \alpha_i x_i(n - t_i). \tag{16}$$

The delay t_i is computed as

$$t_i = \arg \max_{t=1 \cdots T} \langle \tau_t x_i, s \rangle, \forall i \in [1, L]. \tag{17}$$

This approach, although slightly suboptimal compared to the first solution, is much easier to implement. If considering all possible delays within the length of $x_i (T = N)$, the scalar product $\langle \tau_t x_i, s \rangle$ becomes equivalent to the convolution of $s(n)$ and $x_i(n)$

$$\langle \tau_t x_i, s \rangle = x_i * s \tag{18}$$

then (17) can be rewritten as

$$t_i = \arg \max \{x_i * s\}. \tag{19}$$

We can now define an *aligned* database $\tilde{\mathcal{D}} = \{\tau_{t_i} x_i\}_{i=1,\ldots,L}$ and adopt the procedure described in (14) with the modified basis

$$\boldsymbol{\alpha} = (\tilde{D}\tilde{D}^T)^{-1} \tilde{D} s. \tag{20}$$

Fig. 2 shows a segment $s(n)$ containing the notes E3, G3, and C4 (MIDI numbers 52, 55, and 60, respectively).[2] A database D, in Fig. 2(b) and an aligned database $\tilde{D}$, in Fig. 2(c) are used for pitch estimation. It can be seen that results using $\tilde{D}$ [Fig. 2(e)] are more accurate than using the nonaligned database [in Fig. 2(d)]. Indeed, taking phases into account is only

[2]$\mathrm{MIDI}(f_0) = |69 + 12 \cdot \log_2(f_0/440)|$. The first note of the piano $A0 = 27.5$ Hz corresponds to MIDI number 21.

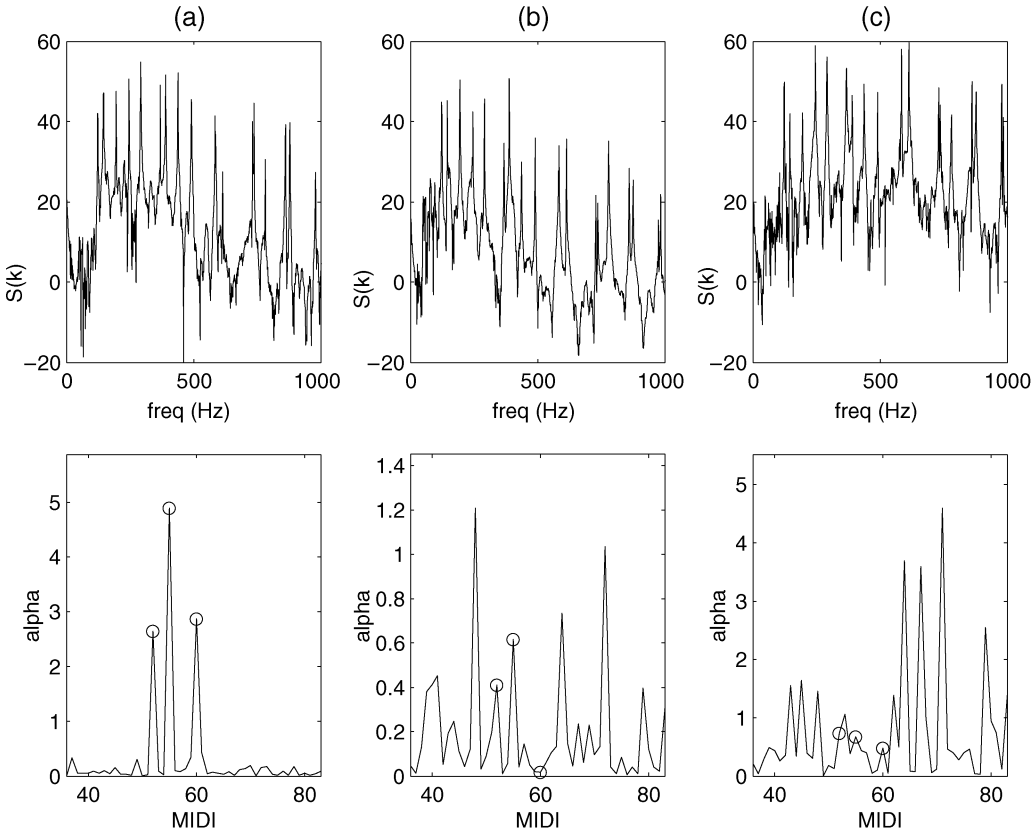

Fig. 3. Calculation of $\boldsymbol{\alpha}$ for equal C-major chords from three different pianos. Top plots: chord spectra. Bottom plots: estimation of $\boldsymbol{\alpha}$ using the same note database D, corresponding to the first piano (circles indicate the actual notes).

beneficial when interferences are constructive, i.e., when signal and database notes are phase-synchronous.

C. Results With a Fixed Database

To validate our method, we have tested our estimation process on instantaneous mixtures of known waveforms, with small random time shifts. The used waveforms belong to three acoustic pianos (to be referred as pianos 1, 2, and 3) from the McGill University's catalog of orchestral sounds [18]. For our experiments, the database D is built with notes from Piano 1.[3]

When tested with samples from the same piano, f_0s are correctly identified even in the presence of harmonic intervals and polyphonies of more than four notes. An example is shown in Fig. 3(a), where the notes from a C major chord are correctly estimated.

However, when tested with samples from pianos 2 and 3, the estimation simply does not work [see Figs. 3(b) and (c) respectively]. In fact, at the top of Fig. 3, it can be noted that the distribution of energy across partials varies considerably between pianos (or recording conditions), thus precluding the use of a simple piano model to estimate notes produced by all pianos. This specificity is critical, as for most real musical recordings we would not have access to a database that completely matches the played sound. Thus, a useful f_0-estimation method must deal with recordings for which such information is not available.

[3]a Steinway and Sons Model D, see info at http://www.steinway.com.

As a possible solution to this problem, we adopt the use of an adaptive approach, where the database of individual notes is estimated for each song using our frequency–domain method. The generated database is then used for time–domain transcription as explained above. In the following sections, we will describe this solution in detail.

D. Adaptive Database Estimation

Estimating the database from the signal $s(n)$ is a three-step process:

1) estimation of "very likely" notes using the frequency–domain method;
2) synthesis of the estimated sounds;
3) interpolation of missing sounds in the database.

1) Note Estimation in the Frequency–Domain: For the method in Section II, let us define ϵ_m as the estimation error rate related to false negatives (FNs) and ϵ_f as the estimation error rate related to false positives (FPs).[4] We can also define ν as the set of parameters that control the transcription process. A usual choice of ν is one that roughly balances the FNs and the FPs, for instance $\nu = \arg\min\{\epsilon_m + \epsilon_f\}$.

However, estimation of all notes in the signal is no longer the goal of our system. The goal is to accurately retrieve a set of "very likely" notes in the signal, i.e., notes that can be detected with a high level of confidence, such that they can be used as the individual waveforms x_i that compose the database D. In

[4]FN are actual notes that are not detected by the system, while FP are nonexisting notes that are detected by the system.

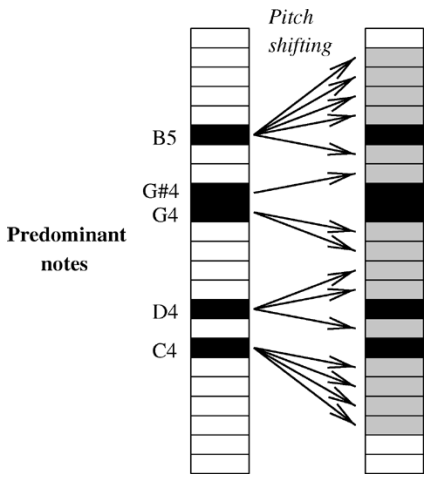

Fig. 4. Gaps in the database are filled by pitch-shifting the estimated notes.

order to do this, we need to retune the algorithm such that its rate of false positives is kept very low. In this paper, ν was set manually after a number of experiments. However, this process could be automated and the parameters learned from a training set of annotated music. Finally, if several occurrences of the same note are found, the one with the highest note-to-signal energy ratio is used, thus favoring the use of notes found in isolation.

2) Synthesis of the Estimated Notes: For the synthesis process on each frame, we use the harmonic combs (Section II-B) of estimated notes to isolate the relevant signal components. We create a modified spectrum by preserving the magnitude and unwrapped phase of these components only. For the synthesis process, we use overlap-add techniques similar to those used in [19].

3) Filling the Gaps: Synthesized notes are organized in the database according to their f_0. The resulting database is incomplete, i.e., does not contain waveforms for all notes in the f_0 range to be estimated; therefore, precluding the use of the time–domain method. The completeness of the database varies depending on the signal and on the parameter set ν. The situation, illustrated on the left-hand side of Fig. 4, leaves us with a database of a few detected notes and many gaps.

A solution is to use a pitch-shifting algorithm to fill the gaps in the database. An efficient pitch-shifting algorithm can be implemented using phase-vocoder theory [16]: we calculate the phase difference $\Delta\varphi$ between consecutive FFT bins in time. We calculate a modified phase difference $\Delta\psi$ as the product between $\Delta\varphi$ and a transposition factor ρ_φ, such that

$$\rho_\varphi = \frac{\tilde{f}_0}{f_0} \tag{21}$$

where f_0 is the original fundamental frequency, and $\tilde{f}_0$ is the fundamental frequency of the shifted note. The modified phase increments are used to synthesise the shifted signals, thus completing the database D as illustrated in the right-hand side of Fig. 4. Shifting is allowed only up to half an octave from the original pitch, to avoid introducing into the database waveforms which are not representative of the piano sound at a given frequency.

IV. Integrating Frame Estimations Over Time

If notes are to be constructed from f_0 estimates, the frame-by-frame analysis needs to be complemented by the analysis of information along the time axis. This applies to estimations using both presented methods.

Let us define a frame estimation as $H_f(f_0, \tau_f)$, where τ_f is the frame time, and f_0 is the estimated frequency. Let us also define a note H_η as the group of frame estimations with fundamental frequency f_0 that occur within a segment defined by an onset time τ_i and offset time τ_e, such that $H_f(f_0, \tau_f) \in H_\eta(f_0, \tau_i, \tau_e)$, if $\tau_f \in [\tau_i, \tau_e]$ and $\Delta\tau_f$, the difference between τ_f and the position of the previous frame estimation, is less than a static predetermined value δ_g. This δ_g value defines the maximum gap that can exist between neighboring frame estimations that belong to the same note. In this implementation, for a hop size of 10 ms, $\delta_g = 30$ ms. Note that the note boundaries τ_i and τ_e are defined by estimations separated by $\Delta\tau_f \geq \delta_g$.

The duration of H_η is evaluated against a predefined threshold δ_d (=40 ms), such that

$$\tau_e - \tau_i = \begin{cases} \geq \delta_d, & \text{then } H_\eta(f_0, \tau_i, \tau_e) \text{ is kept} \\ < \delta_d, & \text{then } H_\eta(f_0, \tau_i, \tau_e) \text{ is eliminated.} \end{cases}$$

If H_η is kept, it is aligned in time to the closest note onset position, calculated using the derivative of the log-energy of the note. This process is equally applied to frequency and time–domain estimations. A more detailed explanation can be found in [10].

V. Results and Discussion

A. About the Evaluation

1) Method: To quantitatively evaluate the accuracy of the transcription, we chose a note by note comparison against a score-like representation of recorded signals. For the evaluation to be meaningful, we require our test signals to be polyphonic, generated by a real piano (i.e., not synthesized) and recorded in live conditions (i.e., presenting the effects of room acoustics). For these signals, we also require score-like representations matching recorded events in time and frequency.

2) Test-Set: We use a collection of MIDI files as our symbolic database. They were generated from performances of amateur and professional piano players and contain a total of 4258 notes. The collection corresponds to segments of piano pieces by five well-known composers: Wolfgang Amadeus Mozart, Ludwig van Beethoven, Claude Debussy, Scott Joplin, and Maurice Ravel. The selection of composers and musical pieces was arbitrary. We use these files to drive a MIDI-controlled acoustic grand piano.[5] The piano was recorded at 44.1-kHz sampling rate, in stereo by using a coupled pair of condenser microphones in a recording studio. The recordings, converted to pulse-code modulation mono wave files at 22 050-Hz sampling rate, are used as input signals for the analysis. The length of the analysis window is 200 ms and overlapping frames are separated by a 10-ms hop.

[5]a Yamaha Disklavier, see info at http://www.yamaha.com/disklavier.

TABLE I
TRANSCRIPTION RESULTS USING THE FREQUENCY-DOMAIN AND TIME-DOMAIN APPROACHES

	FD		TD	
Composer	% TP	% FP	% TP	% FP
Mozart	72.78	14.53	82.20	19.44
Beethoven	65.99	15.42	75.16	21.49
Debussy	74.05	15.90	82.32	20.22
Joplin	60.77	18.59	64.59	29.07
Ravel	68.83	27.40	72.29	33.47
TOTAL	68.98	16.40	76.96	22.33

3) Limitations of Using MIDI Data: The use of MIDI files as our ground truth for evaluation is not without complications: MIDI information contains a list of commands that regulates when to hit and release each piano key, as well as the loudness of the hit. This ignores other features in the sound such as the instrument's sustain, the room's reverberation, the temporal changes in timbre due to the physics of the instrument and the noise produced by the instrument mechanism (e.g., by using the pedals, by the hammering of the strings, etc). It also ignores the mechanically induced delay between note on/off and the real time at which the piano keys are pressed/released. As a consequence, events in the MIDI file will differ from those in the audio signal, most noticeably in the duration of notes and, thus, the polyphony in the signal, and less noticeably in the exact start times of events. Our evaluation tries to compensate for this by using a tolerance window in time when comparing the original and the estimated MIDI files: true positives (TPs) are acknowledged when notes of the same pitch start in both MIDI files within 50 ms of each other. Offset times are not taken into account.

B. Discussion

1) Overall Results: Table I compares results for the transcription using the frequency–domain (FD) and the time–domain (TD) approaches.

For each composer and approach, Table I shows the percentage of true positives (TP), or estimated true notes, and the percentage of false positives (FP), or estimated false notes. For the frequency–domain approach, the overall rate of true positives is almost 70% of the total amount of notes. For the time–domain approach, overall rates of TP increase by 8%, up to almost 77%. This is a significant improvement, considering that as the number of true positives increase, only the most difficult note combinations remain undetected (e.g., harmonically related notes in complex polyphonies). However, the number of FP also increases significantly, raising concerns about the advantages of using this approach. In the following, we will discuss these results in more detail.

2) Polyphony and Harmonicity: For both approaches, the best results are for Debussy's piano pieces, while the worst are for Joplin's segments. This is consistent with the known limitations of transcription systems regarding polyphony and harmonicity: Debussy's segments are based on chromatic melodic progressions, usually with low polyphonies involved, while Joplin's segments are rich in complex polyphonies of highly harmonic sounds, as rag time music usually is. Similar observations can be made for the whole collection. Table II shows the mean, deviation, and maximum polyphonies of the test collection by composer. Polyphony is defined as the number of notes that are *on* when a new note starts. It is measured by counting the number of active notes at every note onset in the MIDI file. We can see the correspondence between the low-order polyphony statistics in the MIDI files and the accuracy of the detection: pieces by Ravel and Joplin show the highest polyphony while producing the lowest detection rates. Pieces by Debussy and Mozart contain lower polyphonies, thus producing more accurate detections. The impact of the polyphony is made worse by the speed of the piece: fast pieces with high polyphonies will provide less steady data for the analysis, thus misleading the estimation process. The last column of Table II shows the mean and deviation of note durations (in seconds) by composer. This data, while strengthening the case for poor estimation in Joplin's music, also partly explains the low number of TP obtained for Beethoven's music (at least with the frequency–domain approach).

TABLE II
TEST COLLECTION STATISTICS FOR POLYPHONY AND NOTE DURATIONS IN SECONDS: AVERAGE (μ), STANDARD DEVIATION (σ), AND MAXIMUM VALUE (MAX)

Composer	polyphony: $\mu \pm \sigma$ (max)	duration $\mu \pm \sigma$
Mozart	1.71 $\pm$ 0.59 (4)	0.41 $\pm$ 0.37
Beethoven	2.09 $\pm$ 1.17 (8)	0.28 $\pm$ 0.26
Debussy	1.99 $\pm$ 1.35 (7)	0.29 $\pm$ 0.33
Joplin	3.56 $\pm$ 1.54 (7)	0.15 $\pm$ 0.17
Ravel	3.48 $\pm$ 1.78 (8)	0.46 $\pm$ 0.56

TABLE III
CATEGORIZATION OF FALSE NEGATIVES (FN) AND FALSE POSITIVES (FP) ACCORDING TO HARMONICITY IN THE MUSIC COLLECTION

Frequency-domain approach

	FN		FP	
Composer	8^{ve}	$3^{rd}/5^{th}$	8^{ve}	$3^{rd}/5^{th}$
Mozart	3.99 %	9.88 %	6.38 %	2.39 %
Beethoven	6.83 %	7.15 %	8.66 %	2.29 %
Debussy	4.19 %	7.63 %	7.08 %	3.47 %
Joplin	19.61 %	3.83 %	7.27 %	3.2 %
Ravel	7.36 %	3.89 %	12.79 %	4.57 %

Time-domain approach

	FN		FP	
Composer	8^{ve}	$3^{rd}/5^{th}$	8^{ve}	$3^{rd}/5^{th}$
Mozart	2.18 %	5.81 %	5.84 %	5.17 %
Beethoven	4.42 %	4.81 %	8.68 %	5.43 %
Debussy	3.18 %	3.95 %	7.89 %	6.78 %
Joplin	17.38 %	3.19 %	9.11 %	8.23 %
Ravel	4.76 %	3.21 %	13.14 %	9.57 %

However, the quantity of notes in the polyphony is only part of the problem. As discussed before, the f_0 relationships between notes have great influence in the accuracy of the detection. Chords with a high harmonic content are more difficult to estimate than those without. Table III shows the contribution of octaves, thirds and, fifths to the estimation error. In the case of false negatives, numbers in the table refer to the percentage of all notes in the MIDI file that were not detected due to harmonic relationships. In the case of false positives, values refer to the percentage of detected notes that were false due to harmonic relationships. All results are categorized by composer.

The problem of harmonicity accounts for more than half the false negatives and around 70% of false positives. Octave intervals are of great significance to miss-detections in Ravel's and

Joplin's music. In Mozart's and Debussy's pieces, it is mostly the presence of thirds and fifths that affects the estimation.

3) Database Estimation: Table III also shows how, in the case of the number of false negatives, the time–domain approach seems to improve upon the frequency–domain method. The percentage of harmonically related FN is consistently reduced by the time–domain method, in some cases for as many as 4% of the total amount of FN. This is evidence to support the idea that if a waveform is successfully incorporated into the basis, then estimations of that note are improved along the duration of the signal. Unfortunately, this is a double-sided argument: if a corrupted waveform (a false positive or a true positive resynthesized from a complex signal context) makes it to the dictionary, then spurious detections are increased. This is reflected in the higher number of FP estimated using the time–domain approach.

Furthermore, even when notes are correctly selected for the database, they do not necessarily represent all instances of the same note during the musical piece: waveforms of the same note are subject to significant variations depending on the energy at which they have been played, their length, the work of the pedals, the context (the interaction with other notes being played at the same time), all of which is neglected by the temporal approach.

What the results show is that the estimation of notes using the time–domain method depends on a successful estimation in the frequency–domain (at the database building stage). Note, for example in Table I, how the increase in the number of TP and FP using the time–domain approach is related to the FP produced by the frequency–domain method: the lower the number of FP in the frequency–domain, the higher the increase in accuracy in the time–domain (e.g., Mozart, Beethoven, and Debussy). On the other hand, less accuracy in the frequency–domain means higher FP rates and lower increments in TP using the time–domain approach. There is a minimum level of accuracy that we need to achieve when building the database to make the time–domain approach work. The assumptions at the core of the method depend on this.

Despite all this, overall results are improved by using the time–domain method, supporting the idea of an alternative approach to the standard analysis in the frequency domain. Moreover, the biggest limitation of the time–domain approach seems to be its dependency on the frequency–domain approach in order to acquire the necessary knowledge for detection. As the theory showed, and the known-database results exhibited, the capabilities of the system are very high given the reliability of the database.

4) Comparison to Other Approaches: There are a number of issues that make a fair comparison of existing approaches to automatic transcription an almost impossible task.

- Databases: There is a lack of standard databases for training and evaluation. Researchers choose the style, instrumentation and acoustic characteristics of the test music as a function of their particular application and/or their access to ground-truth information.
- Annotations: The difficulties of producing reliable ground-truth data usually translates in the use of "shortcuts" to annotation, e.g., synthesized music, midi-driven instruments, score-following, etc. Hand-marking is a painful and time-consuming task that leaves no room for the cross-validation of annotations. Furthermore, there are no standard rules regarding annotations, so different ground-truths are not compatible.
- Evaluation: Different studies use different evaluation methods making numeric comparisons futile, e.g., different criteria for hits or misses, different tolerance windows, etc.

As a reference, we can cite numeric results published on databases of acoustically recorded piano music. Raphael [6], reports rates of 61% TP and 26% FP on a database of 1360 notes from a single piano recording; Carreras *et al.* [7] report average results of 74% TP and 11.7% FP on five segments of piano music (it is important to note that their evaluation assumes perfect onset detection which is not always the case); recently, Marolt [8] reported results of 80.9% TP and 14.7% FP on a database of 3382 notes from three piano recordings. It is worth mentioning that, with the exception of Raphael's approach that uses the Baum–Welch algorithm to train his HMM on the analysis signal, these approaches train their systems prior to analysis on a separate database of isolated notes or tones. By gaining our knowledge directly from the signal, we are not constrained by the limitations imposed by the training set.

VI. Conclusion

In this paper, we concentrate on the problem of automatically transcribing piano music. First, a method is proposed that analyzes the signal on a frame-by-frame basis, detecting the most prominent spectral peaks and grouping them according to expected comb-patterns by means of a set of heuristic rules. The method successfully identifies nearly 70% of notes in acoustic piano recordings. Errors are often associated with the common limitations of frequency–domain approaches: high polyphonies, short durations, and harmonic intervals.

Alternatively, a novel method is presented that identifies notes from polyphonic mixtures in the time–domain. It improves results when facing common issues that arise when using frequency–domain transcription methods. The approach assumes short segments of the original waveform to be the linear sum of weighted individual waveforms (corresponding to the individual notes of the played instrument), and phase relationships to be reproducible. The theory is developed to lead to the conclusion that, by estimating the values of the mixing vector $\boldsymbol{\alpha}$ for each frame, accurate polyphonic pitch detection can be achieved. This is true provided that we have the original waveform and that we have a database of waveforms corresponding to individual notes of the instrument. To this end, two conditions need to be satisfied, phase-alignment, obtained through independent shifting of each vector, and a reliable database, constructed by using the results of the frequency–domain approach. This method has the advantage of not needing prior training on a separate database of isolated notes. Moreover, the use of the dual approach could be further refined to use results of the time–domain method to improve on the frequency–domain estimation, thus increasing the accuracy of the detection with each iteration.

The time–domain method improves estimations by 8%, but with a cost to the reliability of the detection of nearly 6%. Detailed analysis concludes that a considerable improvement to the previous approach can be obtained when the rate of false positives from the frequency–domain is reasonably low, i.e., when a certain level of accuracy has been reached. Otherwise, false positives are introduced into the the database, badly affecting the performance of the new method.

In the future, results could be improved by adding knowledge about the physical behavior of the instrument into our model, as proposed by [9] (inharmonicity, string coupling, etc). Also, for the time–domain method, we could go beyond our crude initial approximation that the waveforms are invariant with respect to loudness, except by a global scaling factor, and add a model for nonlinear timbre variations, such as the ones used in piano synthesis [20].

The generalization of this approach to other instruments depends on sources being static, as any movement will, at the very least, destroy phase relationships of wavelengths lower than the amplitude of the movement. Also, the phase relationships need to be reproducible in the excitation, a condition that constrains the nature of possible sources to a few instruments (e.g., keyboards: organ, harpsichord).

ACKNOWLEDGMENT

The authors would like to thank the S2M team at the Laboratoire de Mécanique et d'Acoustique, Marseille, France, for kindly letting them use their Yamaha Diskclavier.

REFERENCES

[1] J. A. Moorer, "On the transcription of musical sound by computer," *Comput. Music J.*, vol. 1, no. 4, pp. 32–38, 1977.

[2] M. Piszczalski and B. A. Galler, "Automatic music transcription," *Comput. Music J.*, vol. 1, no. 4, pp. 24–31, 1977.

[3] A. Klapuri, T. Virtanen, and J. M. Holm, "Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals," in *Proc. COST-G6 Conf. Digital Audio Effects (DAFx-00)*, Verona, Italy, 2000, pp. 141–146.

[4] K. D. Martin, A Blackboard System for Automatic Transcription of Simple Polyphonic Music MIT Media Lab, Perceptual Computing Section, Tech. Rep. 385, Jul. 1996 [Online]. Available: ftp://sound.media.mit.edu/pub/Papers/kdm-TR385.ps.gz

[5] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka, "Application of Bayesian probability network to music scene analysis," in *Computational Auditory Scene Analysis*, D. F. Rosenthal and H. G. Okuno, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, May 1998.

[6] C. Raphael, "Automatic transcription of piano music," in *Proc. 3rd Int. Conf. Music Inf. Retrieval*, Paris, France, Oct. 2002, pp. 15–19.

[7] F. Carreras, M. Leman, and M. Lesaffre, "Automatic harmonic description of musical signals using schema- based chord decomposition," *J. New Music Res.*, vol. 28, no. 4, pp. 310–333, 1999.

[8] M. Marolt, "Networks of adaptive oscillators for partial tracking and transcription of music recordings," *J. New Music Res.*, vol. 33, no. 1, pp. 49–59, 2004.

[9] L. I. Ortiz-Berenguer, F. J. Casajús-Quirós-Quiros, and S. Torres-Guijarro, "Multiple piano note identification using a spectral matching method with derived patterns," *J. Audio Eng. Soc.*, vol. 53, no. 1/2, pp. 32–43, Jan./Feb. 2005.

[10] J. P. Bello, "Toward the automated analysis of simple polyphonic music: A knowledge-based approach," Ph.D. dissertation, Univ. London, [Online]. Available: http://www.elec.qmul.ac.uk/staffinfo/juan/

[11] A. Klapuri, "Number theoretical means of resolving a mixture of several harmonic sounds," in *Proc. Eur. Signal Process. Conf.*, 1998, pp. 2365–2368.

[12] M. Marolt, "On finding melodic lines in audio recordings," in *Proc. 7th Int. Conf. Digital Audio Effects*, Naples, Italy, 2004, pp. 217–221.

[13] B. Mulgrew, P. Grant, and J. Thompson, *Digital Signal Processing: Concepts and Applications*. Hampshire, U.K.: Palgrave Macmillan, 1998.

[14] P. Lepain, "Polyphonic pitch extraction from musical signals," *J. New Music Res.*, vol. 28, no. 4, pp. 296–309, 1999.

[15] M. Dolson, "The phase vocoder: a tutorial," *Comput. Music J.*, vol. 10, no. 4, p. 14, 1986.

[16] D. Arfib, F. Keiler, and U. Zölzer, "Time–frequency processing," in *Digital Audio Effects*. New York: Wiley, 2002.

[17] N. H. Fletcher and T. D. Rossing, *Phys. Musical Instruments*. New York: Springer Verlag, 1991.

[18] F. Opolko and J. Warpnick, McGill University Master Samples (MUMS), Faculty of Music, McGill Univ., Montreal, QC, Canada CD-ROM set, 1989.

[19] A. De Gotzen, N. Bernardini, and D. Arfib, "Traditional (?) implementations of a phase vocoder: the tricks of the trade," in *Proc. COST G-6 Conf. Digital Audio Effects*, Verona, Italy, Dec. 2000, pp. 37–44.

[20] J. Bensa, F. Gibaudan, K. Jensen, and R. Kronland-Martinet, "Note and hammer velocity dependence of a piano string model based on coupled digital waveguides," in *Proc. Int. Comput. Music Conf.*, La Habana, Cuba, 2001.

Juan P. Bello (M'06) was born in 1976. He received the engineering degree in electronics from the Universidad Simon Bolivar, Caracas, Venezuela, in 1998 and the Ph.D. degree on the automatic transcription of simple polyphonic music from Queen Mary, University of London, London, U.K., in 2003.

Currently, he is the Technical Manager of the Centre for Digital Music at Queen Mary, University of London. His research is mainly focused on the semantic analysis of musical signals and its applications to music information retrieval, digital audio effects, and music interaction.

Laurent Daudet (M'02) received the degree in statistical and nonlinear physics from the Ecole Normale Supérieure, Paris, France, in 1997 and the Ph.D. degree in mathematical modeling from the Université de Provence, Marseilles, France, on audio coding and physical modeling of piano strings in 2000.

In 2001 and 2002, he was a Marie Curie Postdoctoral fellow at the Department of Electronic Engineering at Queen Mary, University of London, London, U.K. Since 2002, he has been working as Lecturer at the Université Pierre et Marie Curie (Paris 6), Paris, France, where he joined the Laboratoire d'Acoustique Musicale. His research interests include audio coding, time-frequency and time-scale transforms, sparse representations of audio, and music signal analysis.

Mark B. Sandler (M'87-SM'95) was born in London, U.K., in 1955. He received the B.Sc. and Ph.D. degrees from University of Essex, Essex, U.K., in 1978 and 1984, respectively.

He is Professor of Signal Processing at Queen Mary University of London, London, U.K., where he moved in 2001 after 19 years at King's College London. For 18 months, he was founder and CEO of Insonify Ltd., an Internet Audio Streaming startup. He has published over 250 papers in journals and conferences.

Prof. Sandler is a Fellow of the IEE and a Fellow of the Audio Engineering Society. He is a two-time recipient of the IEE A. H. Reeves Premium Prize.