Automatic Transcription of Guitar Chords and Fingering From Audio

Ana M. Barbancho, Anssi Klapuri, *Member, IEEE*, Lorenzo J. Tardón, and Isabel Barbancho, *Senior Member, IEEE*

***Abstract*—This paper proposes a method for extracting the fingering configurations automatically from a recorded guitar performance. 330 different fingering configurations are considered, corresponding to different versions of the major, minor, major 7th, and minor 7th chords played on the guitar fretboard. The method is formulated as a hidden Markov model, where the hidden states correspond to the different fingering configurations and the observed acoustic features are obtained from a multiple fundamental frequency estimator that measures the salience of a range of candidate note pitches within individual time frames. Transitions between consecutive fingerings are constrained by a musical model trained on a database of chord sequences, and a heuristic cost function that measures the physical difficulty of moving from one configuration of finger positions to another. The method was evaluated on recordings from the acoustic, electric, and the Spanish guitar and clearly outperformed a non-guitar-specific reference chord transcription method despite the fact that the number of chords considered here is significantly larger.**

***Index Terms*—Acoustic signal analysis, chord transcription, hidden Markov model (HMM), multiple fundamental frequency estimation, music signal processing.**

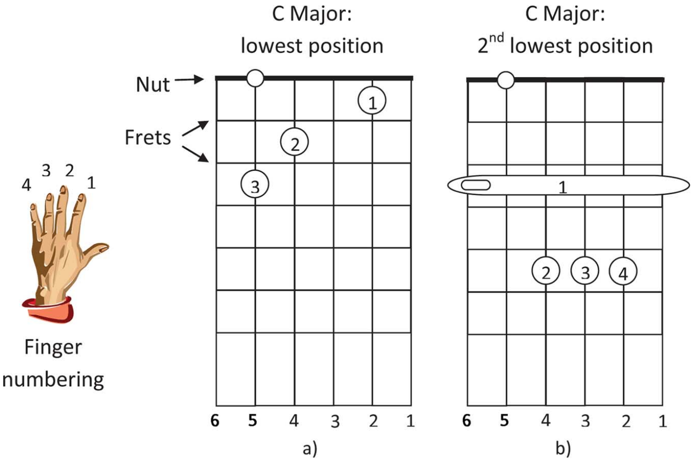

Fig. 1. Example chord charts showing two alternative fingerings for the C major chord. The vertical lines indicate the six strings and the circled numbers on the strings indicate the positions of the fingers. Figure (b) shows a barre chord where the index finger presses down several strings simultaneously.

I. Introduction

THE GUITAR is one of the most popular instruments in Western music, being extensively used in genres such as blues, country, flamenco, jazz, pop, and rock. The instrument is portable, affordable, enables a high degree of musical expression, and allows an intuitive notation in the form of tablatures and chord charts (see Fig. 1 for an example). For these reasons, the guitar is the instrument of choice of many professional and amateur musicians alike—the latter often learning to play it independently, without a teacher. The guitar comes in many forms, the main variants being the acoustic and the electric guitar. Right-handed people generally use the right hand to excite the strings by plucking or strumming and the left hand to choose the note pitches by pressing down the strings at appropriate positions along the guitar neck.

In this paper, a guitar-specific automatic chord transcription system is presented. Automatic chord transcription from musical audio in general has been the subject of a number of research papers over the past few years. These have been motivated by the applications of music retrieval [1], such as audio-to-score synchronization [2], and automatic lead sheet generation [3]. Most methods consist of two parts: extraction of acoustic features to describe the tonal content of the signal, and a search algorithm to find the most likely sequence of chords by matching the features with pre-stored internal models. Most methods use the so-called *chroma* vectors as acoustic features, where the spectrum of a signal is estimated with the resolution of 12 bins per octave and the corresponding bins in different octaves are then collapsed into a single 12-dimensional feature vector [4], [5]. This representation is motivated by the human perception of harmony and was first termed *pitch class profile* [6], [7], although the term chroma has been later widely adopted. Chord recognition systems have introduced various improvements to the simple chroma features, such as the estimation of the reference tuning [8], [9], spectral whitening to improve their timbre-invariance [5], suppression of percussive instruments, and feature decorrelation [10]. Some methods have taken advantage of the additional information conveyed by the bass note by introducing separate chroma vectors for the bass and the treble range [3], [11]. These methods were based on approximate transcription of the note pitches in music, instead of just spectrum estimation [3], [11].

The first chord transcription algorithms were based on matching the chroma features against 12-dimensional binary templates representing different chords [7], [8]. More recently, statistical methods have been employed, most notably hidden Markov models (HMMs) [9], [11]–[14], but also conditional random fields [15] and dynamic Bayesian networks [3] have been used. Some systems take advantage of the statistical dependencies between chords and other musical features by

Manuscript received April 25, 2011; accepted October 14, 2011. Date of publication October 31, 2011; date of current version January 25, 2012. This work was supported by the Ministerio de Ciencia e Innovación of the Spanish Government under Project TIN2010-21089-C03-02. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. James Johnston.

A. M. Barbancho, L. J. Tardón, and I. Barbancho are with the Department Ingeniería de Comunicaciones, E.T.S. Ingeniería de Telecomunicación, Universidad de Málaga, 29071 Málaga, Spain (e-mail: abp@ic.uma.es; lorenzo@ic.uma.es; ibp@ic.uma.es).

A. Klapuri is with the Centre for Digital Music, Queen Mary University of London, London E1 4NS, U.K. (e-mail: anssi.klapuri@elec.qmul.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2174227

estimating jointly the chords and the musical key [16], [17], the chords and the downbeat positions [14], or the chords, bass notes, downbeat positions, and the key [3]. Comparative evaluations of chord transcription methods can be found in [15], [18] and in the annual MIREX[1] evaluation.

Automatic extraction of chords and fingering from recorded guitar performances has several applications, including automatic accompaniment systems, musically oriented computer games, score typesetting systems, and instrument tutors that give feedback for a student's playing. However, none of the chord transcription methods mentioned above has been specifically developed for the guitar, although much of guitar music is based on switching between a limited number of discrete chord and fingering configurations. There are several methods for estimating the plucking point location of an individual guitar string played in isolation [19]–[21], and very few methods doing this for all the strings played simultaneously [22]. Some guitar transcription systems exist that are based on video data without audio. Most of these, while accurate, are obtrusive to the guitarist: cameras must be mounted on the guitar [23] or the guitarist must wear colored fingertips to be tracked [24]. A few approaches can be found that use both audio and video to identify the guitar chords of a performance [25], [26]. In [27], the authors propose a neural-network based system to estimate the number of sounding strings in a guitar chord and use this information to propose fingerings that can be used to produce the chord, without explicitly detecting the fingering applied on the recording. In [28], a system for *hexaphonic* guitar transcription is presented, meaning that the sound of each of the six guitar strings is captured separately using a special-purpose multichannel microphone installed on the guitar.

In this paper, we propose a method for analyzing audio recordings of guitar performances. The method identifies not only the chords played, but also the fingering configuration on the guitar fretboard. As illustrated in Fig. 1, a given musical chord can be played using a number of alternative fingerings, and this leads to a large number of fingering configurations to be recognized [29]. In the proposed system, 330 different fingering configurations are considered, which is significantly more than in previous audio-based chord transcription systems.[2] At the same time, a high degree of accuracy is achieved, since the method utilizes acoustic and musical models specifically developed and trained for the guitar. In particular, we propose a cost function for estimating the physical difficulty of switching from one fingering configuration to another and incorporate that into our model. The method is limited to solo recordings where only the guitar is playing, without other instruments.

The rest of this paper is organized as follows. Section II describes the proposed method in detail, including the features and the models employed. Section III presents simulation results for the proposed method and a reference method and Section IV summarizes the main conclusions.

II. Proposed Method

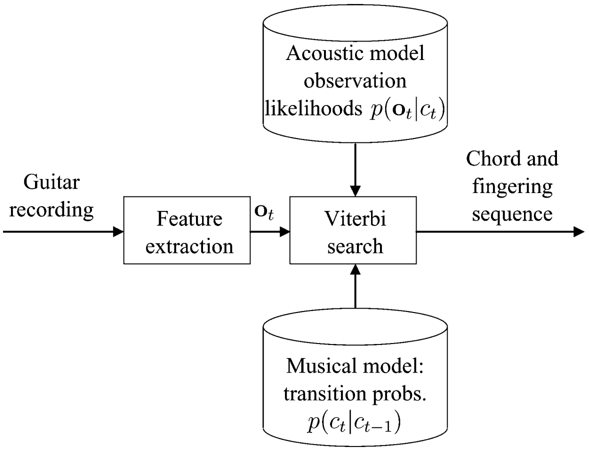
Acoustic model observation likelihoods $p(\mathbf{o}_t|c_t)$
Guitar recording
Feature extraction
$\mathbf{o}_t$
Viterbi search
Chord and fingering sequence
Musical model: transition probs. $p(c_t|c_{t-1})$

Fig. 2. Overview of the proposed system.

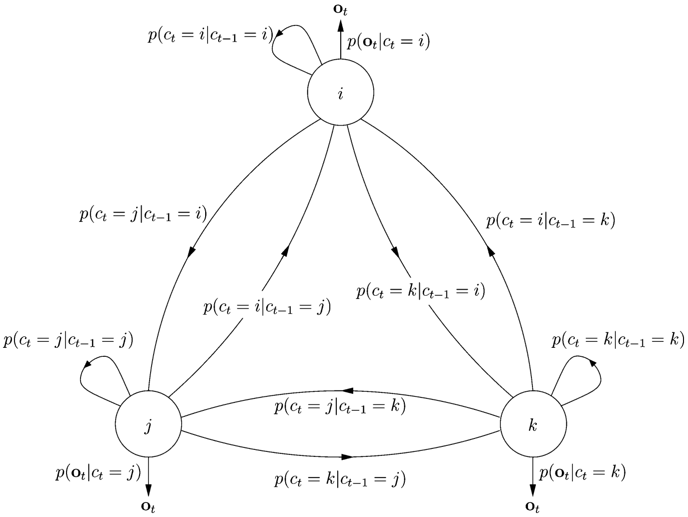

Fig. 3. Illustration of the employed HMM, with only three hidden states shown. Each hidden state corresponds to a unique CFC c. The observation vectors $\mathbf{o}_t$ are obtained using a multiple-F0 estimator.

Fig. 2 shows an overview of the proposed method. A recorded waveform of a guitar performance is processed with a multiple fundamental frequency (F0) estimator [30] which measures the salience (strength) of different F0 candidates in each time frame t. The observation vector $\mathbf{o}_t$ consists of the estimated saliences of 41 discrete note pitches between 82 Hz (E2) and 830 Hz (G#5). The search for an optimal sequence of chord and fingering configurations (CFCs) is constrained by two pretrained models: an acoustic model describing the probability $\mathrm{p}(\mathbf{o}_t \mid c_t)$ of observing $\mathbf{o}_t$ given a certain CFC, and a musicological model that determines the probability $\mathrm{p}(c_t \mid c_{t-1})$ of switching between temporally successive CFCs.

We formulate our method as a fully connected HMM [31], [32] where the hidden states correspond to the different CFCs. Fig. 3 illustrates the model, however showing only three states for practical reasons. The state-conditional observation densities $\mathrm{p}(\mathbf{o}_t \mid c_t)$ are learned from annotated training data. The transition probabilities $\mathrm{p}(c_t \mid c_{t-1})$ are obtained by combining statistics of chord sequences in manually annotated training data with a heuristic model representing the physical difficulty of switching between two fingering configurations.

The number of hidden states in the HMM, $N = 330$, is determined by the amount of unique CFCs considered. We consider only the four most common chord types (major, minor, major 7th, and minor 7th) at the 12 different root positions $(\mathrm{C}, \#\mathrm{C}, \mathrm{D}, \#\mathrm{D}, \ldots, \mathrm{B})$, however including several different variants of each chord. Fig. 4 illustrates 21 basic fingering configurations that form the basis for the 330 unique CFCs. In 12 out

[1] http://www.music-ir.org/mirex/ (date last viewed 22.10.2011)

[2] In previous chord transcription systems, the number of chords considered has varied from 24 [11], [14], [17], [16], [10] to 36 [13] to 48 [8] to 109 [3] and 147 [12].

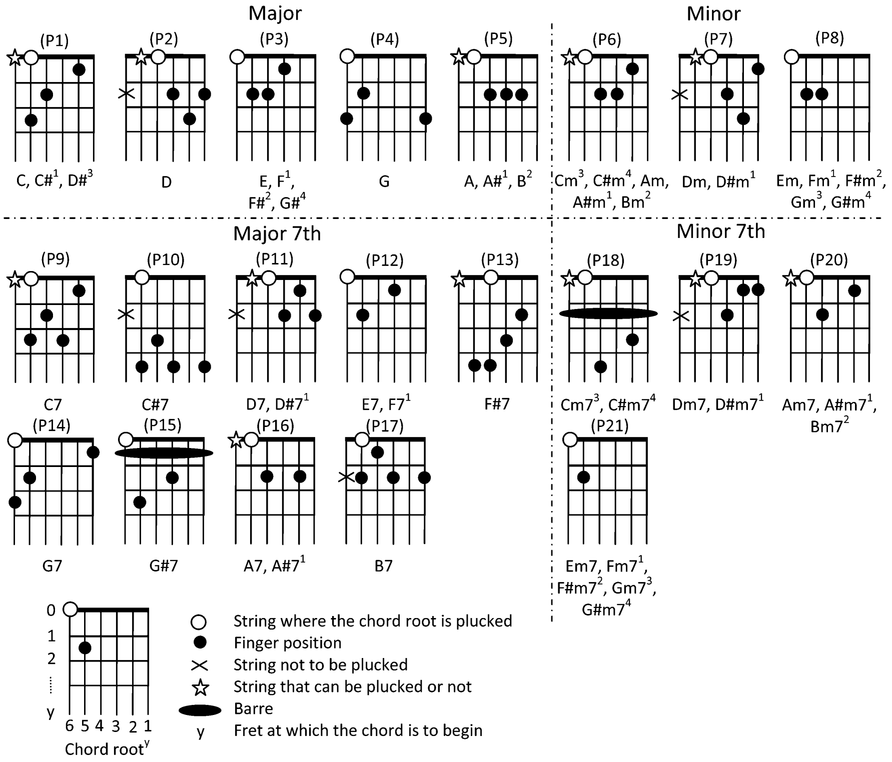

Fig. 4. Chord charts for the 21 basic fingerings. See the guide at the bottom of the figure for the interpretation of the chord charts.

of these 21 basic configurations, the number of plucked strings (four, five, or six) can vary, leading to 33 configurations when the number of active strings is taken into account. Note that in Fig. 4, the strings that can be either plucked or not in a given chord are marked with a star. Finally, each of the 33 configurations can be played at different positions along the guitar neck by using a *capo* or just the fingers. To cover the full length of the guitar neck, we consider ten different positions (frets $0, 1, \ldots, 9$) for each of the 33 basic configurations, leading to $N = 330$ unique CFCs and hidden states in the HMM. The selected hand positions are the most common ones for guitar music. However, it is noteworthy that more states could be defined adding more hand positions.

A. Feature Extraction

Instead of relying on the chroma features discussed in Section I, the front-end feature extraction of our system is based on the multiple-F0 estimator presented in [30]. The method calculates the salience of a given F0 as a weighted sum of the amplitudes of its harmonic partials in a spectrally whitened signal frame. The output of the estimator, $s_{t,\tau}$, represents the salience of the fundamental period τ in frame t under analysis. For convenience, the fundamental frequency candidates are expressed as unrounded MIDI[3] note numbers

$$F(\tau) = 69 + 12\log_2((f_s/\tau)/440) \tag{1}$$

where $f_s = 44\,100$ Hz is the sampling rate and 69 is an agreed-upon number for the reference pitch 440 Hz.

The entire vector of pitch saliences $s_{t,\tau}$ is not retained, but instead, an observation vector is constructed which contains the saliences of 41 discrete pitch values between $n_{\mathsf{min}} = 40$ and $n_{\mathsf{max}} = 80$. These are the MIDI numbers of the lowest (E2) and the highest note (G#5) considered, respectively. The salience of note n is obtained by selecting the maximum salience within ± 0.5 semitone range around the note n:

$$s_t(n) = \max_i \; s_{t,i}, \quad i \in \{\tau || F(\tau) - n| \leq 0.5\} \tag{2}$$

The observation vector $\mathbf{o}_t$ is then given by

$$\mathbf{o}_t = [s_t(n_{\mathsf{min}}), s_t(n_{\mathsf{min}} + 1), \ldots, s_t(n_{\mathsf{max}})]^{\mathsf{T}}. \tag{3}$$

B. State-Conditional Observation Likelihoods

Annotated training data of recorded guitar performances were used to estimate the probability density function (pdf) of the observations. The training data was annotated semi-automatically by utilizing a multichannel microphone (based on Roland GK-3 guitar microphone) to record the sound of each string separately. The "ground truth" finger position on a given string was then estimated from the waveform of the individual string using the YIN algorithm [34]. The total (polyphonic) sound of the guitar was recorded with an external microphone placed at approximately 0.5 m distance from the sound hole of the acoustic guitar, and in the case of the electric guitar, by utilizing the signal obtained by connecting a cable to the electric guitar directly in a normal manner. The polyphonic signal was used to extract the observation vectors for both the train and the test data.

Two models were trained. The first pdf, $\mathsf{p}(s(n)\,|\,n)$, describes the observed salience values (2) given that note n was played at that time according to the annotation of the training material. Similarly, the pdf $\mathsf{p}(s(n)\,|\,\bar{n})$ describes the observed salience values when note n was not played according to the annotation (here we omit time index for clarity, since the pdfs are independent of time t). The two pdfs are approximated with normal and lognormal distributions, respectively,

$$\mathsf{p}(s(n)\,|\,n) = \frac{1}{\sigma_{s\,|\,n}\sqrt{2\pi}} \exp\left[-\frac{(s(n) - \mu_{s\,|\,n})^2}{2\sigma^2_{s\,|\,n}}\right] \tag{4}$$

[3] Musical instrument digital interface (MIDI) is a standard for exchanging data between electronic musical devices [33].

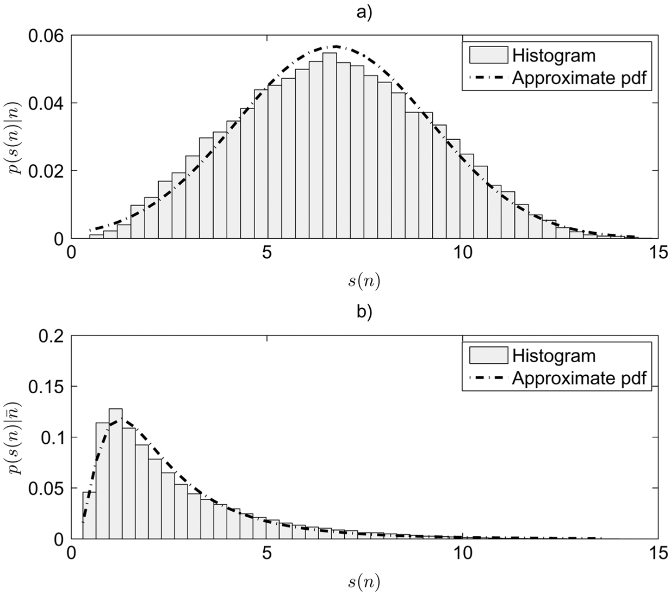

Fig. 5. (a) The bars show sample histogram of the salience values $s(n)$ in the training data at times when note n was played according to the annotation. The curve shows the corresponding model pdf $p(s(n)\,|\,n)$. (b) Sample histogram of the salience values $s(n)$ in the training data when the note n was not played according to the annotation (bars), and the corresponding model pdf (curve).

$$\mathsf{p}(s(n)\,|\,\bar{n}) = \frac{1}{s(n)\sigma_{s\,|\,\bar{n}}\sqrt{2\pi}} \exp\left[-\frac{(\ln s(n) - \mu_{s\,|\,\bar{n}})^2}{2\sigma^2_{s\,|\,\bar{n}}}\right] \tag{5}$$

where $\mu_{s\,|\,n}$ and $\sigma_{s\,|\,n}$ are, respectively, the mean and the standard deviation of the salience observation when note n is actually played, and similarly, $\mu_{s\,|\,\bar{n}}$ and $\sigma_{s\,|\,\bar{n}}$ are the mean and the standard deviation of the log-salience when note n is *not* played (although other notes may be active). We estimated these parameters from the training data using the unbiased estimators [35].

The two pdfs are illustrated in the upper and lower panels of Fig. 5, respectively, together with the sample histograms of the salience values in the training data. For convenience in the following, we use $f_n(s(n)) \equiv \mathsf{p}(s(n)\,|\,n)$ and $f_{\bar{n}}(s(n)) \equiv \mathsf{p}(s(n)\,|\,\bar{n})$ as shorthands for these two pdfs.

Each chord and fingering configuration c is characterized by the notes that are played on each string or—for the strings that should not be plucked—the notes that could be mistakenly played on the string. As a result, each CFC combination (state c of the HMM) is described by two vectors

$$\mathbf{n}_c = [n_c(1), n_c(2), \ldots, n_c(6)]^{\mathsf{T}}$$
$$\mathbf{q}_c = [q_c(1), q_c(2), \ldots, q_c(6)]^{\mathsf{T}}$$

where the vector $\mathbf{n}_c$ contains the MIDI number of the notes $n_c(\,\cdot\,) \in \{n_{\mathsf{min}}, \ldots, n_{\mathsf{max}}\}$ that may be played on each of the six strings, and the vector $\mathbf{q}_c(\,\cdot\,)$ contains binary values $q_c \in \{0, 1\}$ indicating if each string is to be plucked or not. For strings that are not to be plucked, $\mathbf{n}_c$ indicates the lowest note that could potentially be mistakenly played on that string, given the positions of the other fingers.

The state-conditional observation likelihoods $\mathsf{p}(\mathbf{o}_t\,|\,c_t)$ are defined with the help of the vectors $\mathbf{n}_c$ and $\mathbf{q}_c$, under the assumption that the salience observations corresponding to the six strings are independent of each other given $\mathbf{n}_c$ and $\mathbf{q}_c$:

$$\mathsf{p}(\mathbf{o}_t\,|\,c_t) = \prod_{i=1}^{6} \mathsf{p}(s_t(n_c(i))\,|\,n_c(i), q_c(i)) \tag{6}$$

The probabilities on the right-hand side of (6) are obtained from the two pdfs (4)–(5) introduced above. If a string is to be plucked $(q_c(i) = 1)$:

$$\mathsf{p}(s_t(n_c(i))\,|\,n_c(i), q_c(i) = 1) = f_n(s_t(n_c(i))) \tag{7}$$

where $f_n(s(n)) \equiv \mathsf{p}(s(n)\,|\,n)$ is given by (4).

When the string is not to be plucked $(q_c(i) = 0)$, we require that no note is played on string i within the range of pitch values that the performer could physically reach given the finger positions $\mathbf{n}_c$ for the active (plucked) strings. This requirement can be implemented as follows. We first find the maximum salience in range $n_c(i), \ldots, n_c(i) + \Delta$:

$$\bar{s}_t(n_c(i)) = \max_{j=0,\ldots,\Delta} s_t(n_c(i) + j) \tag{8}$$

where $n_c(i)$ now defines the lowest finger position that the player can reach on the not-to-be-plucked string and $\Delta = 3$ determines the range of reachable positions above that. Then, the probabilities needed in (6) for the case $q_c(i) = 0$ are given by

$$\mathsf{p}(s_t(n_c(i))\,|\,n_c(i), q_c(i) = 0) = f_{\bar{n}}(\bar{s}_t(n_c(i))) \tag{9}$$

where $f_{\bar{n}}(s(n)) \equiv \mathsf{p}(s(n)\,|\,\bar{n})$ is given by (5).

C. Transition Probabilities Between CFCs

The transition probabilities $\mathsf{p}(c_t\,|\,c_{t-1})$ represent the probability of switching between two temporally consecutive CFCs. This information can be stored in matrix $\mathbf{A}$, where each entry

$$a_{ij} = \mathsf{p}(c_t = j\,|\,c_{t-1} = i). \tag{10}$$

The transition probabilities $\mathsf{p}(c_t\,|\,c_{t-1})$ that we use consist of two factors, one representing the musical probability of switching between any two chords and the other representing the physical difficulty of moving fingers from one configuration to another:

$$\mathsf{p}(c_t\,|\,c_{t-1}) \propto \mathsf{p}_{\mathsf{mus}}(c_t\,|\,c_{t-1}) \cdot \mathsf{p}_{\mathsf{phy}}(c_t\,|\,c_{t-1}). \tag{11}$$

The musical transition probabilities $\mathsf{p}_{\mathsf{mus}}(c_t\,|\,c_{t-1})$ were estimated from the manual chord annotation for the first eight albums of The Beatles, as provided by Harte *et al.* [8]. The database also includes complex chords, such as augmented and diminished chords, which were left out from the training data. The musical transitions probabilities are independent of the key: only the chord type and the distance (0–11 semitones) between the chord roots matters. For example, a transition from G major 7th chord to C major chord is regarded as a transition from a major 7th chord to a major chord with distance 5 semitones. This means that the musical transition probabilities are defined between a set of $4 \times 12 = 48$ different chords and that the musical transition probabilities are identical for many of the 330 CFCs. Note that, from the point of view of the composition of western music, the transition probabilities between different chords does not depend on the instrument [36].

The other factor in (11), $\mathsf{p}_{\mathsf{phy}}(c_t\,|\,c_{t-1})$, represents the physical difficulty of switching between two fingering configurations. The "cost" of moving fingers was approximated by adding up simple vertical and horizontal movements of the fingers. A simple vertical movement is considered as passing one finger from a string to the adjacent string. A simple horizontal movement is considered as passing a finger from a fret to the adjacent one. We consider that there is no cost if a finger is removed. The cost of adding a finger is defined as the vertical cost of moving the finger from string 1 to the selected string. In a barre chord, finger 1 is defined to be located on the string 6 at the corresponding fret. The cost due to the different fingers are summed up. For example, the cost of moving

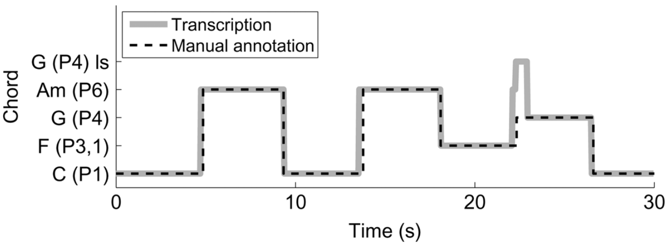

Fig. 6. Chord transcription and the manual annotation of the chord progression C Am F G found in many popular music songs, here played on the acoustic guitar. The numbers in the parentheses indicate the fingering (see text for details).

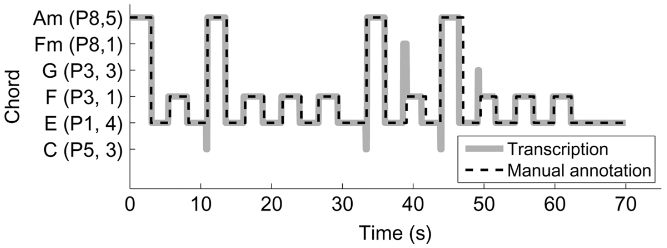

Fig. 7. Chord transcription and the manual annotation of "Sin documentos" by Los Rodriguez played on the Spanish guitar. The numbers in the parentheses indicate the fingering (see text for details).

between the two CFCs shown in Fig. 1 is $4+2$ (vertical and horizontal cost of finger 1) $+3$ (finger 2) $+2+2$ (finger 3) $+1$ (adding finger 4) which adds up to the total cost of 14. Note that following these ideas, the physical difficulty factor $\mathsf{p}_{\mathsf{phy}}(c_t \mid c_{t-1})$ could also be tuned for specific genres or users. The physical transition probability $\mathsf{p}_{\mathsf{phy}}(c_t \mid c_{t-1})$ is then defined as

$$\mathsf{p}_{\mathsf{phy}}(c_t = j \mid c_{t-1} = i) \propto \frac{1}{b_{ij}} \qquad (12)$$

where b_{ij} denotes the cumulated cost of moving fingers from configuration i to configuration j. Although there is no theoretical motivation for the precise form of (12), the obtained transition probability values for different chords correspond reasonably well with experience on the difficulty of finger movements, and this form worked well in practice.

The transition probabilities given by the musical and the physical model are combined according to (11). The resulting probability values are smoothed using the Witten–Bell algorithm [37]. Finally, three free parameters remain that control the amount of chord changes and were set experimentally. These are the probability of staying in the same CFC, the transition probabilities between different fingerings of the same chord, and the transition probabilities within the same fingering but with different number of active (plucked) strings.

D. Finding the Most Likely State Sequence

After the model parameters have been trained, chord transcription is carried out by finding the most likely HMM state sequence $c_{1:T} = c_1 c_2 \ldots c_T$ given a sequence of observation vectors $\mathbf{o}_{1:T} = \mathbf{o}_1 \mathbf{o}_2 \ldots \mathbf{o}_T$:

$$c^*_{1:T} = \arg\max_{c_{1:T}} \left[\mathsf{p}(\mathbf{o}_1 \mid c_1) \prod_{t=2}^{T} \mathsf{p}(\mathbf{o}_t \mid c_t) \mathsf{p}(c_t \mid c_{t-1}) \right]. \qquad (13)$$

The state sequence is found using the Viterbi algorithm [38], [31]. Initial probabilities $\mathsf{p}(c_1)$ of the states are assumed to be uniform and are therefore omitted from (13).

Fig. 6 illustrates the results for a simple chord progression on the acoustic guitar and Fig. 7 shows the results for "Sin documentos" by Los Rodriguez played with the Spanish guitar. In these figures, the four chord types are indicated by X (major), Xm (minor), X7 (major 7th), and Xm7 (minor 7th), where X is replaced by the chord root $(\mathrm{C}, \#\mathrm{C}, \ldots, \mathrm{B})$. The two numbers $(\mathrm{P}n, m)$ in the parentheses indicate one of the basic fingerings Pn shown in Fig. 4 and the fret m where the fingering should start. For example E(P2,2) represents the E major chord played with the fingering $P2$, shown in Fig. 4 with fingering, starting after fret 2. The text "ls" is placed after the label in the case that the chord is played by plucking smaller number of strings than usually. The method does not detect silent segments but outputs a chord label in each frame.

The most likely state sequence found by maximizing (13) is post-processed to remove any detected chords that are shorter than three frames in duration. These short "chords" were found to be due to the transition noise between consecutive chords. Accordingly, the onset time of the chord following the deleted chord is made slightly earlier.

III. Results

The proposed transcription method was quantitatively evaluated using recorded guitar performances to be described below. The results are compared with those obtained using the method proposed earlier by Ryynänen and Klapuri [11]. Note that this and other methods for chord detection—like those in the MIREX evaluation, are not designed to obtain fingering and plucking information.

Separate data sets were used for training and testing the proposed method. The training data set consisted of 22 recordings: 8 recordings of an acoustic guitar and 14 recordings of an electrical guitar played by two different guitar players, totalling 2 hours and 39 minutes in duration. The test data consisted of 14 recordings: 11 recordings of two different Spanish guitars played by two different guitar players, 2 recordings of an electric guitar and 1 recording of an acoustic guitar, totalling 21 minutes and 50 seconds in duration. It should be noted that the training data contains recordings of guitars with steel strings only, however the test data contains recordings with both steel and nylon strings. Acoustic guitars were recorded using one microphone placed in front of the guitar sound hole at about 0.5 m distance. The signal from an electric guitar was obtained via a cable plucked into the guitar directly, without applying any distortion or other effects. The test data was annotated manually by listening through the audio recordings and with the help of knowing the fingering sequence that was given to the players to perform.

The chord transcription method is evaluated by comparing the transcribed chords with the manually annotated reference chords frame-by-frame. An individual frame covers 46 ms of the corresponding target recording. The performance metric is defined as the ratio of the number of correctly detected chords to the total number of frames analyzed.

Table I shows the performance of the proposed method ("PM") on the test data in different but related detection tasks. The performance of the proposed method using only the observations and the physical transition probabilities (i.e., $\mathsf{p}(c_t \mid c_{t-1}) \propto \mathsf{p}_{\mathsf{phy}}(c_t \mid c_{t-1})$) is also shown (labeled "PHY"). The performance of the proposed method considering only the musical transition probabilities (i.e., $\mathsf{p}(c_t \mid c_{t-1}) \propto \mathsf{p}_{\mathsf{mus}}(c_t \mid c_{t-1})$) is indicated with "MUS." Finally, the performance of the proposed method without transition probabilities ("PC") is evaluated by using uniform transitions probabilities (i.e., $\mathsf{p}(c_t \mid c_{t-1}) = \mathrm{const.}$). Note that the removal

TABLE I
RESULTS FOR THE PROPOSED METHOD

Correct chord (48 possibilities; fingering not considered)			
PM	PHY	MUS	PC
95%	83%	86%	75%

Correct chord and fingering (210 possibilities, number of plucked strings not considered)	
PM	PHY
88%	79%

Correct chord, fingering and number of plucked strings (330 possibilities)	
PM	PHY
87%	71%

TABLE II
RESULTS FOR THE REFERENCE METHOD [11]

Correct major/minor chord (24 possibilities, 7th chords not considered)	91%
Correct major/minor chord and root correct in 7th chords (24 possibilities: A7 labeled as A regarded as correct)	80%
Correct chord (48 possibilities: all chords considered)	70%

of the physical model in our system means that the fingering and the number of plucked strings cannot be estimated. Observe that the top part of Table I indicates the performance of the system in the process of detecting the correct chord from among 48 possibilities (4 chord types × 12 different chord roots). The middle part shows the detection performance for the correct chord and fingering which can be viewed as selecting a state from among 210 alternatives as explained in the beginning of Section II. Finally, the bottom part considers the detection of the correct chord, fingering and the number of plucked strings, which may vary in some of the configurations. This latter case corresponds to the utilization of the complete model and the identification of a state from among 330 possibilities. Recall that the two simplified models "MUS" and "PC" cannot detect the fingering and the number of plucked strings because they do not take into account the physical difficulty of fingering transitions.

Table I shows that including the physical difficulty model (PHY) and the musical probability (MUS) improve the chord detection performance when compared with the model that does not use the transitions probabilities (PC). Note that the best performance is obtained using the complete model which means that both the physical model and the musical model contain valuable information for the detection task.

To put the results in perspective, we compare the results with those of the general-purpose chord transcription system of Ryynänen and Klapuri [11]. The reference method has been previously shown to achieve results comparable to the state-of-the-art in the MIREX 2008 audio chord detection task.[4] It has been developed for general musical audio recordings that typically consist of a mixture of different instruments and singing voice.

Table II shows the results for the method of Ryynänen and Klapuri [11] on the same test data that was employed to evaluate the proposed system in Table I. The reference method does not include the 7th chords, but only detects major and minor chords and one among the 12 chord roots, therefore the results are presented separately for the cases where the 7th chords are excluded from the test data or where labelling Cm7 as Cm, for example, would be regarded as correct. The best direct comparison between the proposed and the reference method is obtained by comparing the first row of Table I with the second row of Table II.

As can be seen, the results of the proposed method are significantly better than those of the reference method, despite the fact that the proposed method considers a significantly larger set of chord and fingering configurations. The improvement is mostly attributed to the fact that the proposed system was specifically developed and trained for the guitar. This allowed us to employ acoustic models and musical heuristics that are specific to the guitar.

IV. CONCLUSION

A method was presented for the detection of the sequence of chords and fingering configurations from audio recordings of guitar performances. The proposed method labels each frame of the input signal with a certain major, minor, major 7th, or minor 7th chord, along with the corresponding left-hand fingering on the guitar fretboard. The method performed significantly better than the general-purpose chord transcription method [11] used as a reference, despite the fact that the proposed method analyzes the target recordings to a greatly finer detail, choosing one among 330 different chord and fingering configurations. The improvement is attributed to the musical models that were specifically developed for the guitar and for the acoustic models that were trained using acoustic material from the guitar. This shows that developing music transcription systems for more narrowly targeted contexts can lead to significantly improved performance. The accuracy of the proposed system makes it a viable candidate for use in many real-world applications, such as musically oriented computer games, interfaces in score typesetting systems, and automatic accompaniment systems.

[4] http://www.music-ir.org/mirex/wiki/2008:MIREX2008_Results (date last viewed 22.10.2011).

REFERENCES

[1] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proc. IEEE*, vol. 96, no. 4, pp. 668–696, Apr. 2008.

[2] C. Fremerey, M. Müller, and M. Clausen, "Handling repeats and jumps in score-performance synchronization," in *Proc. 11th Int. Soc. Music Inf. Retrieval Conf. (ISMIR-10)*, Utrecht, The Netherlands, 2010, pp. 243–248.

[3] M. Mauch and S. Dixon, "Simultaneous estimation of chords and musical context from audio," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1280–1289, Aug. 2010.

[4] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, 2001, pp. 15–18.

[5] M. Müller and S. Ewert, "Towards timbre-invariant audio features for harmony-based music," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 649–662, Mar. 2010.

[6] C. Krumhansl, *Cognitive Foundations of Musical Pitch*. New York: Oxford Univ. Press, 1990.

[7] T. Fujishima, "Real time chord recognition of musical sound: A system using common lisp music," in *Proc. Int. Comput. Music Conf. (ICMC)*, Beijing, China, 1999, pp. 464–467.

[8] C. Harte, M. Sandler, S. Abdallah, and E. Gómez, "Symbolic representation of musical chords: A proposed syntax for text annotations," in *Proc. 6th Int. Society Music Inf. Retrieval Conf. (ISMIR-05)*, London, U.K., 2005, pp. 66–71.

[9] J. Bello and J. Pickens, "A robust mid-level representation for harmonic content in music signal," in *Proc. 6th Int. Soc. Music Inf. Retrieval Conf. (ISMIR-05)*, London, U.K., 2005, pp. 304–311.

[10] Y. Ueda, Y. Uchiyama, T. Nishimoto, N. Ono, and S. Sagayama, “HMM-based approach for automatic chord detection using refined acoustic features,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, 2010, pp. 5518–5521.

[11] M. Ryynänen and A. Klapuri, “Automatic transcription of melody, bass line and chords in polyphonic music,” *Comput. Music J.*, vol. 32, pp. 72–86, Fall 2008.

[12] A. Sheh and D. P. W. Ellis, “Chord segmentation and recognition using EM-trained hidden Markov models,” in *Proc. 4th Int. Conf. Music Inf. Retrieval (ISMIR-04)*, Baltimore, MD, 2003, pp. 183–189.

[13] K. Lee and M. Slaney, “Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 291–301, Feb. 2008.

[14] H. Papadopoulos and G. Peeters, “"Simultaneous estimation of chord progression and downbeats from an audio file,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, 2008, pp. 121–124.

[15] J. A. Burgoyne, L. Pugin, C. Kereliuk, and I. Fujinaga, “A cross-validated study of modelling strategies for automatic chord recognition in audio,” in *Proc. 8th Int. Soc. Music Inf. Retrieval Conf. (ISMIR-07)*, Vienna, Austria, 2007, pp. 251–254.

[16] T. Rocher, M. Robine, P. Hanna, and L. Oudre, “Concurrent estimation of chords and keys from audio,” in *Proc. 11th Int. Society Music Inf. Retrieval Conf. (ISMIR-10)*, Utrecht, The Netherlands, 2010, pp. 141–146.

[17] J. Pauwels and J.-P. Martens, “Integrating musicological knowledge into a probabilistic framework for chord and key extraction,” in *Proc. 128th Audio Eng. Society Conv. (AES-2010)*, London, U.K., May 2010, pp. 1–9.

[18] H. Papadopoulos and G. Peeters, “Large-scale study of chord estimation algorithms based on chroma representation and HMM,” in *Proc. Int. Workshop Content-Based Multimedia Indexing*, Bordeaux, France, 2007, pp. 53–60.

[19] I. Barbancho, L. J. Tardón, A. M. Barbancho, and S. Sammartino, “Pitch and played string estimation in classical and acoustic guitars,” in *126th Audio Eng. Soc. Conv. (AES-2009)*, Munich, Germany, 2009, pp. 1–9.

[20] H. Penttinen, J. Siiskonen, and V. Välimäki, “Acoustic guitar plucking point estimation in real time,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, 2005, pp. 209–212.

[21] C. Traube and P. Depalle, “Deriving the plucking point location along a guitar string from a least-square estimation of a comb filter delay,” in *Proc. Canadian Conf. Elect. Comput. Eng. (CCECE2003)*, Montreal, QC, Canada, 2003, pp. 2001–2004.

[22] J. Diego and I. Barbancho, “Graphical interface for string and fret estimation in guitar,” in *Proc. 7th Int. Symp. Comput. Music Modeling Retrieval (CMMR2010)*, Málaga, Spain, 2010, pp. 273–274.

[23] A. Burns and M. Wanderley, “Visual methods for the retrieval of guitarist fingering,” in *Proc. Conf. New Interfaces for Musical Expression (NIME’06)*, Paris, France, 2006, pp. 196–199.

[24] C. Kerdvibulvech and H. Saito, “Vision-based guitarist fingering tracking using a Bayesian classifier and particle filters,” in *Proc. Pacific-Rim Symp. Image Video Technol. (PSIVT’07)*, Santigo de Chile, Chile, 2007, pp. 625–638.

[25] A. Hrybyk and Y. Kim, “Combined audio and video analysis for guitar chord identification,” in *Proc. 11th Int. Soc. Music Information Retrieval Conf. (ISMIR-10)*, Utrecht, The Netherlands, 2010, pp. 159–164.

[26] M. Paleari, B. Huet, A. Schutz, and D. Slock, “A multimodal approach to music transcription,” in *Proc. 15th IEEE Int. Conf. Image Process. (ICIP08)*, San Diego, CA, 2008, pp. 93–96.

[27] T. Gagnon, S. Larouche, and R. Lefebvre, “A neural network approach for preclassification in musical chords recognition,” in *Proc. Conf. Rec. 37th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, 2003, pp. 2106–2109.

[28] P. D. O’Grady and S. T. Rickard, “Automatic hexaphonic guitar transcription using non-negative constraints,” in *Proc. IET Irish Signals Syst. Conf. (ISSC2009)*, Dublin, Ireland, 2009, pp. 1–6.

[29] C. López, The “First Stage” Guitar Chord Chart First Stage Concepts, 2006.

[30] A. Klapuri, “Multiple fundamental frequency estimation by summing harmonic amplitudes,” in *Proc. 7th Int. Soc. Music Inf. Retrieval Conf. (ISMIR-06)*, Victoria, BC, Canada, Oct. 2006, pp. 1–6.

[31] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, no. 2, pp. 257–289, Feb. 1989.

[32] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York: Wiley, 2001.

[33] The Complete MIDI 1.0 Detailed Specification. The MIDI Manufacturers Association, 1996 [Online]. Available: www.midi.org, website: (date last viewed 04/07/11)., 2nd. ed.

[34] A. de Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[35] S. M. Kay, “Fundamentals of statistical signal processing,” in *Estimation Theory*, ser. Signal Processing Series. Englewood Cliffs, NJ: Prentice-Hall, 1993, vol. I.

[36] D. Temperley, *The Cognition of Basic Musical Structures*. Cambridge, MA: MIT Press, 2001.

[37] D. Jurafsky and J. Martin, *Speech and Language Processing*. Upper Saddle River, NJ: Prentice-Hall, 2000.

[38] A. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *IEEE Trans. Inf. Theory*, vol. IT-13, no. 2, pp. 260–269, Apr. 1967.

Ana M. Barbancho received the degree in telecommunications engineering and the Ph.D. degree from University of Málaga, Málaga, Spain, in 2000 and 2006, respectively. In 2001, she also received the degree in solfeo teaching from the Málaga Conservatoire of Music.

Since 2000, she has been with the Department of Communications Engineering, University of Málaga, as an Assistant and then Associate Professor. Her research interests include musical acoustics, digital signal processing, new educational methods, and mobile communications.

Dr. Barbancho was awarded with the “Second National University Prize to the Best Scholar 1999/2000” by the Ministry of Education of Spain in 2000 and with the “Extraordinary Ph.D. Thesis Prize” of ETSI Telecomunicación of University of Málaga in 2007.

Anssi Klapuri (M’06) received the Ph.D. degree in information technology from Tampere University of Technology (TUT), Tampere, Finland, in 2004.

He visited as a post-doc researcher at Ecole Centrale de Lille, France, and Cambridge University, U.K., in 2005 and 2006, respectively. He worked until 2009 as a Professor and Group Leader at TUT. In 2009, he joined Queen Mary, University of London, London, U.K., as a Lecturer. His research interests include audio signal processing, auditory modeling, and machine learning. He has worked as a principal investigator in numerous industrial research projects.

Proc. Klapuri received the IEEE Signal Processing Society 2005 Young Author Best Paper Award.

Lorenzo J. Tardón received the degree in telecommunications engineering from University of Valladolid, Valladolid, Spain, in 1995 and the Ph.D. degree from Polytechnic University of Madrid, Madrid, Spain, in 1999.

In 1999 he worked for ISDEFE on air traffic control systems at Madrid-Barajas Airport and for Lucent Microelectronics on systems management. Since November 1999, he has been with the Department of Communications Engineering, University of Málaga, Málaga, Spain. He is currently the head of the Application of Information and Communications Technologies (ATIC) research group. He has worked as main researcher in several projects on music analysis. His research interests include digital signal processing, serious games, audio signal processing, and pattern recognition.

Isabel Barbancho (SM’10) received the degree in telecommunications engineering and the Ph.D. degree from University of Málaga (UMA), Málaga, Spain, in 1993 and 1998, respectively. In 1994, she also received her degree in piano teaching from the Málaga Conservatoire of Music.

Since 1994, she has been with the Department of Communications Engineering ,UMA, as an Assistant and then Associate Professor. She has been the main researcher in several research projects on polyphonic transcription, optical music recognition, music information retrieval and intelligent content management. Her research interests include musical acoustics, signal processing, multimedia applications, audio content analysis and serious games.

Dr. Barbancho received the Severo Ochoa Award in Science and Technology, Ateneo de Málaga-UMA in 2009.