Multipitch Estimation of Piano Music by Exemplar-Based Sparse Representation

Cheng-Te Lee, *Student Member, IEEE*, Yi-Hsuan Yang, *Member, IEEE*, and Homer H. Chen, *Fellow, IEEE*

***Abstract*—Pitch, together with other midlevel music features such as rhythm and timbre, holds the promise of bridging the semantic gap between low-level features and high-level semantics for music understanding. This paper investigates the pitch estimation of a piano music signal by exemplar-based sparse representation. A note exemplar is a segment of a piano note, stored in the dictionary. We first describe how to represent a segment of the piano music signal as a linear combination of a small number of note exemplars from a large note exemplar dictionary and then show how the sparse representation problem can be solved by l_1-regularized minimization. The proposed approach incorporates tuning factor estimation, note candidate selection, and hidden-Markov-model-based smoothing into the estimation process to improve accuracy. Unlike previous approaches, the proposed approach does not require retraining for a new piano. Instead, only a dozen notes of the new piano are needed. This feature is computationally attractive and avoids intense manual labeling. The system performance is evaluated using 70 classical music recordings of two real pianos under different recording conditions. The results show that the proposed system outperforms four state-of-the-art systems.**

***Index Terms*—Content retrieval, l_1-regularized minimization, music transcription, pitch estimation, sparse representation.**

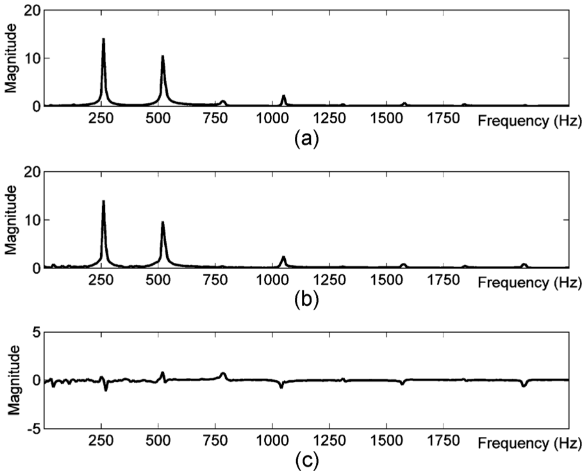


Fig. 1. Illustration of source number ambiguity and octave ambiguity. The spectrum of a single piano note C4 (MIDI note number 60) is almost identical to the spectrum of the mixture of C4 and C5. C5 is one octave apart from C4. Therefore, one cannot tell the number of sources from the spectrum of the music signal. Neither can one tell whether C5 is voiced or not from the spectrum alone, referred to as octave ambiguity. (a) Spectrum of C4. (b) Spectrum of the mixture of C4 and C5. (c) Signal difference between (a) and (b).

I. Introduction

CONTENT-BASED music information retrieval (MIR) has drawn more attention due to the explosive growth of digital music. However, the semantic gap between the high-level human perception and the low-level signal features [1] remains a challenging issue for content-based MIR systems. It is commonly believed that midlevel music features, such as pitch, rhythm, and timbre can help bridge the gap since they are more closely related to the human perception of music [1], [2]. This motivates us to investigate pitch estimation for piano music.

Piano is one of the most popular musical instruments worldwide. A monophonic sound produced by a piano has quasiharmonic spectra [3]. The lowest frequency of the harmonic series is called the pitch or the fundamental frequency of the monophonic sound. Multiple monophonic sounds combined together become a polyphonic sound. Multipitch estimation refers to the determination of the underlying pitches of a polyphonic sound. Unlike monopitch estimation, multipitch estimation has to deal with issues such as the source number ambiguity and the octave ambiguity (see Fig. 1). In this regard, multipitch estimation is more challenging than monopitch estimation [4], [5]. Similar to the work described in [6]–[9], this paper focuses on the multipitch estimation problem of piano music. Although only one single type of musical instrument is considered, the problem is by no means trivial. In fact, Peeters [5] showed that pitch estimation for piano is more challenging than that of any other instruments because pianos have a wide range of pitches and can play multiple notes at the same time.

A significant amount of work has been done to extract multiple pitch contours from a music signal generated by either a single type of instrument or multiple types of instruments [6]–[22]. These methods can be divided into two categories: 1) learning based and 2) dictionary based. Although useful for multipitch estimation, either category has its drawbacks.

Manuscript received October 02, 2011; revised February 28, 2012; accepted March 05, 2012. Date of publication March 20, 2012; current version published May 11, 2012. This work was supported in part by grants from the National Science Council of Taiwan under Contracts NSC 100-2221-E-002-198-MY3 and NSC 100-2218-E-001-009 and in part by a grant from the National Taiwan University under Contract 10R80919-5. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mrinal Mandal.

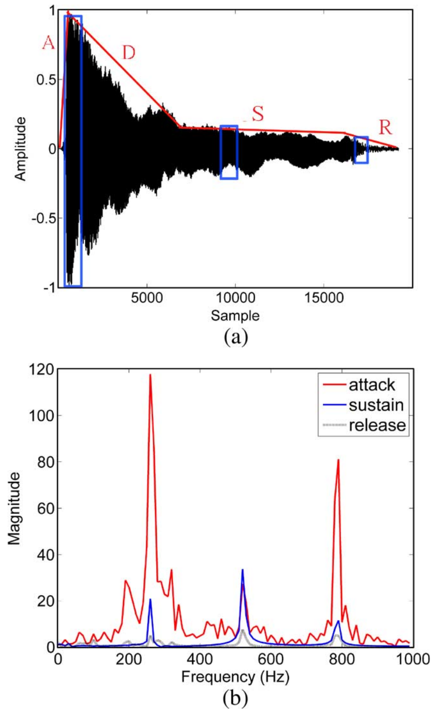
C.-T. Lee is with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: aderleee@gmail.com).

Y.-H. Yang is with the Research Center for IT Innovation, Academia Sinica, Taipei 11529, Taiwan (e-mail:yang@citi.sinica.edu.tw).

H. H. Chen is with the Department of Electrical Engineering, Graduate Institute of Communication Engineering, and Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 10617, Taiwan (e-mail: homer@cc.ee.ntu.edu.tw).

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/TMM.2012.2191398



(a)

(b)

Fig. 2. (a) The waveform and its corresponding ADSR (attack, decay, sustain, and release) envelope of note C4 played by a piano. Each selected frame is marked by a box. (b) The spectra of the three selected frames.

The learning-based methods [6]–[14], when dealing with instruments of various brands, require acoustic training data (which are random chords or excerpts of musical pieces) and the corresponding MIDI files [23] as the ground truth to learn the spectral characteristics of musical sounds. Collecting such training data is time-consuming [24]. On the other hand, the dictionary-based methods [15]–[21] attempt to describe a note by a single representative frame-level spectral shape (often called template). However, as shown in Fig. 2, the shape of the short-time spectrum of a note changes with time. Such spectral variation characterizes a sound [25]. Therefore, overlooking the spectral variation will not deliver the best possible performance for multipitch estimation.

To overcome the aforementioned drawbacks, we propose a novel approach that is free of learning and takes the ADSR envelope into consideration. The idea is somewhat similar to the dictionary-based methods described in [18], [20], [21], but we store a number of segments of a note [e.g., those frames depicted in Fig. 2(a)] rather than the note template in the dictionary. Therefore, each note in our approach corresponds to more than one atom in the dictionary, and each atom represents a particular segment of the note. Here, each atom is referred to as a *note exemplar*. In this way, updating our dictionary for a new piano can be easily done by adding the note exemplars of the new piano into the dictionary. In addition, only a dozen sample notes of the new piano are needed. In other words, our approach does not require retraining. In contrast, traditional learning-based approaches require a new set of training data and, hence, a model retraining whenever a new piano is considered.

To enhance the accuracy of multipitch estimation, we impose the so-called *sparsity* constraint on the number of active pitches. Specifically, each short-time representation of an input piano signal is expressed as a linear combination of the short-time representations of a small subset of the note exemplar dictionary. The adoption of the sparsity constraint is based on the observation that only a few pitches are active as a music piece unfolds. Fig. 3 illustrates the idea of the exemplar-based sparse representation approach, where the short-time representation of a signal within a frame is the magnitude spectrum of the framed signal.

Exemplar-based approaches employing the sparsity constraint have been proposed for object classification [26]–[28]. However, little effort, if any, along this vein has been made for multipitch estimation of piano music [29], [30].

To evaluate the pitch estimation accuracy of our approach, an extensive performance study is conducted on the recordings of two real pianos (cf. Section V). Magnitude spectrum coefficients of the music signals are found to produce better accuracy than other low-level features, such as MFCCs, MDCT coefficients, and time-domain signal. The accuracy improves as the number of exemplars increases, indicating that an overcomplete dictionary is favorable [26].

The primary advantages of our approach are as follows.

- Acoustic training data and the corresponding ground truth MIDI files are not needed.
- Learning of pitch templates or dictionaries is not required.
- Adapting to a new piano is much easier, requiring only a dozen sample notes.
- A more realistic modeling of the musical sound is achieved by taking the ADSR envelope into account.
- Significant accuracy improvement over the state-of-the-art approaches.

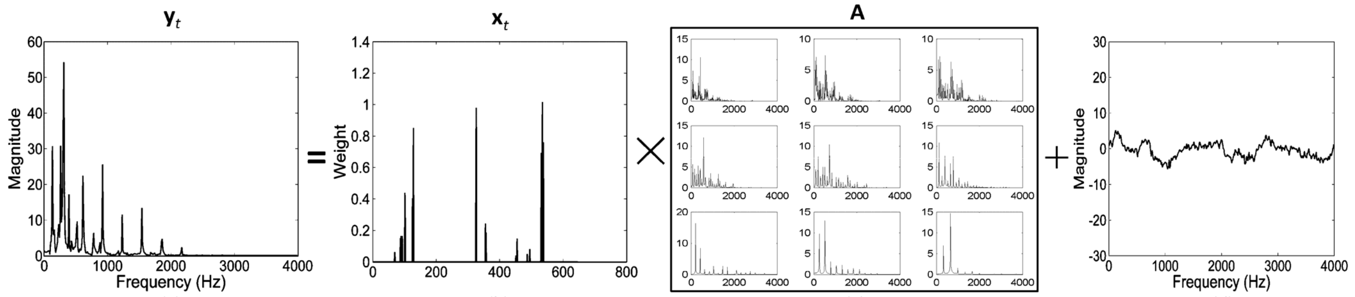
The organization of this paper is as follows. Section II reviews the related work, and Section III introduces the exemplar-based approach. In Section IV, we describe the proposed multipitch estimation system in detail. Section V describes an evaluation of performance of the proposed system on 70 recordings of solo piano music with respect to previous systems. Section VI concludes this paper.

II. Related Work[1]

Monopitch estimation, compared to multipitch estimation, is relatively easy to deal with as described in Section I. Therefore, we focus on polyphonic music and discuss related methods.

Moorer [22] made the first attempt to extract multipitch from polyphonic music. Although limited to duets, it has inspired many interesting subsequent works. As mentioned earlier, there are two categories of approaches to multipitch estimation: learning based and dictionary based. In this section, we review the basic principle of these approaches. In addition, we review sparse representation [26], which is a useful method for enhancing the performance of multipitch estimation.

[1] The main differences between the conference and journal version are: • More detailed descriptions and more empirical analysis of each component of the proposed system are provided (cf. Sections III, IV and V-D). • A new temporal smoothing method, which is based on the coupled HMM rather than the conventional HMM, is proposed, (cf. Sections IV-C and V-B). • The number of recordings in the test data set is increased from 10 to 70 to provide a thorough evaluation of the proposed system. (cf. Section V).



(a) (b) (c) (d)

Fig. 3. Illustration of the exemplar-based sparse representation of frame spectrum for multipitch estimation. A frame spectrum is represented as a linear combination of a small number of spectra of note exemplars from a large dictionary plus the residue. (a) The magnitude spectrum of a frame. (b) The coefficients of sparse representation. (c) The spectra of note exemplars. (d) The residue.

A. Learning-Based Approaches

The basic idea of learning-based approaches is to learn the temporal, spectral, or spectrotemporal relationship between the music signals in the training data set and the underlying pitches (obtained from the corresponding MIDI file) and use the resulting relationship to estimate the pitches of a music signal in the test data set [31]. This relationship is related to the harmonic structure or the smoothness of the music spectrum. The harmonic structure refers to the number of harmonics of a musical sound and the ratio of their magnitudes. It is found that harmonic structure is unique to each musical instrument and is a robust feature for sound source separation [32]. Various approaches have been proposed for harmonic structure modeling. Some consider the whole spectrum while others take an additional step to detect the peaks (local maxima) of the spectrum and use the resulting information for multipitch estimation. For example, neural networks [6] and support vector machines (SVMs) [8] are applied to model the harmonic structure of individual piano notes based on the whole spectrum. In contrast, the rule-based approach [7] is based on clusters of spectral peaks, and the maximum-likelihood approach [14] uses the peaks as well as the nonpeak regions of the spectrum for multipitch estimation.

The smoothness of spectra is another important characteristic of musical sounds useful for multipitch estimation because nearby harmonics tend to have similar magnitude [32]. This particular characteristic is often used as an additional constraint to refine multipitch estimation [9]–[12].

There are two common drawbacks of the learning-based approaches. First, their estimation accuracy degrades significantly for an unfamiliar sound with spectral characteristics drastically different from those of the training signals [8]. This is illustrated in Fig. 4, where we see that different pianos or different recording conditions result in different spectral shapes for a note. Second, updating the model for a new instrument is difficult because it would require additional MIDI scripts of the instrument, which are not always available. In addition, the model retraining process can be time-consuming.

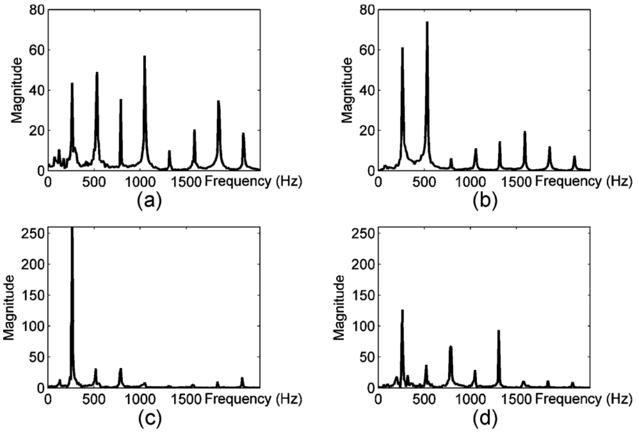


Fig. 4. Spectrum of C4 produced by (a) a Yamaha Disklavier piano using ambient recording, (b) the same Yamaha Diskalavier piano using close recording, (c) a Steinway D piano using ambient recording, and (d) the same Steinway D piano using close recording. As we can see, different pianos produce sounds with different spectral shapes, and the recording condition affects the spectral shape of piano sounds.

B. Dictionary-Based Approaches

Dictionary-based approaches aim at decomposing the magnitude spectrum or other short-time representations of an input music signal into a weighted sum of atoms [33]. Each atom carries certain information related to pitch or instrument.

Given a sequence of observation vectors $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T]$, standard components analysis methods, such as independent components analysis (ICA) and non-negative matrix factorization (NMF), can be applied in an unsupervised fashion to decompose $\mathbf{Y}$ into a matrix of atoms $\mathbf{A}$ and a coefficient matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$, where $\mathbf{y}_t = \mathbf{A}\mathbf{x}_t$ [33]–[35]. However, ICA is not applicable when one uses features, such as magnitude spectra, for short-time representation as it may produce negative values in $\mathbf{A}$ that contradict the physical meaning of a spectrum. NMF is free from this problem because it explicitly constrains $\mathbf{A}$ and $\mathbf{X}$ to be non-negative [15], [16], [19], under the assumption that the atoms are purely additive. However, due to the nature of unsupervised learning, identities of atoms are not certain [33]. Thus, manual labeling of pitch of each atom is required.

Approaches have been proposed that enforce each atom to represent only one note, either by an additional supervised learning process using sample notes [18], [20] or by initializing the atoms to specific values using domain knowledge [21]. However, such hybrid approaches suffer from the common weakness of learning-based approaches described before.

Since only a small number of notes are simultaneously played as a musical piece unfolds, a feasible approach is to constrain $\mathbf{x}_t$ to be sparse. That is, a coefficient vector with a small number of nonzero entries [16]–[18] is favored. Early attempts impose a sparsity constraint on an overdetermined system [18], [20], [36].

C. Sparse Representation Method

Recent years have witnessed a growing interest in the application of sparse representation in communication and computer science. Sparse representation describes an input signal as a linear combination of a few atoms (base elements) from a large dictionary. The method has been shown to be effective for audio classification [37] and coding [38]. Audio classification is performed by treating the coefficients of sparse representation as audio features, and data compression is achieved by encoding a small number of nonzero coefficients.

A recent development in the theory of sparse representation and compressed sensing [39], [40] suggests that the minimal l_1-norm solution is also the sparsest solution in most times, given that the dictionary is overcomplete. It is reported in the literature [26] and empirically validated in our evaluation (cf. Section V) that an underdetermined system (which uses an overcomplete dictionary) is preferable to an overdetermined one.

Motivated by the aforementioned observation, Wright *et al.* proposed the first exemplar-based sparse classification method that directly uses all training data as atoms to build an overcomplete dictionary for robust face recognition [26]. This exemplar-based method for dictionary construction has also been shown useful for audio tasks, such as music genre classification [27] and automatic speech recognition [28]. In these systems, each atom is associated with a predefined class.

The extension of this exemplar-based method to multipitch estimation, however, is nontrivial since it is typical to have multiple pitches be simultaneously present in an input signal. The assumption that only one class of object (face, genre, or phoneme) is present in the input signal may not hold. Therefore, in addition to directly using note exemplars as atoms, we also employ tuning factor estimation and note candidate selection as front-end and hidden Markov Model (HMM)-based smoothing as back end to enhance our mutipitch estimation system.

III. Exemplar-Based Sparse Representation

In this section, we describe how to extract the pitch information of piano music with the exemplar-based sparse representation method. We first describe the basic idea underlying this method, and then go into the details of how to construct a dictionary of note exemplars. Finally, we describe a straightforward implementation of the exemplar-based sparse representation method and point out its pitfalls. The findings collected from this pilot study lead to the development of three additional components that complete the proposed system.

A. Frame-by-Frame Multipitch Estimation

Our method divides input music into overlapping short-time frames and performs pitch estimation for each frame individually. The basic idea behind this method is to represent the short-time representation of a frame as a linear combination of the short-time representations of note exemplars. Then, a sparsity constraint is imposed to favor the linear combination with a minimal number of nonzero terms. The problem can be formulated as the following l_1-minimization problem:

$$\widehat{\mathbf{x}}_t = \arg\min \|\widehat{\mathbf{x}}_t\|_1 \text{ subject to } \mathbf{y}_t = \mathbf{A}\mathbf{x}_t \qquad (1)$$

where $\mathbf{y}_t$ is a short-time representation vector of the tth input music frame, $\mathbf{A} = [\mathbf{A}, \mathbf{A}_2, \ldots, \mathbf{A}_K]$ is the note exemplar dictionary, K is the number of notes, $\mathbf{A}_i = [\mathbf{a}_{i,1}, \mathbf{a}_{i,2}, \ldots, \mathbf{a}_{i,Ni}]$ is a collection of short-time representations of the note exemplars of the ith note of piano. When each note is represented by the same amount of exemplars (i.e., $N_i = D$), the total number of note exemplars N in $\mathbf{A}$ is equal to KD. Since the same feature representation is used for the input music frame and the note exemplars, the dimensions $\mathbf{y}_t$ and $\mathbf{a}_{i,n}$ are equal.

To take the effect of noise into consideration, we introduce a small positive real number ε to allow minor error in the decomposition and reformulate the problem as

$$\widehat{\mathbf{x}}_t = \arg\min \|\widehat{\mathbf{x}}_t\|_1 \text{ subject to } \|\mathbf{y}_t - \mathbf{A}\mathbf{x}_t\|_2^2 \leq \varepsilon. \qquad (2)$$

The minimization problem is equivalent to

$$\widehat{\mathbf{x}}_t = \arg\min \left(\|\mathbf{y}_t - \mathbf{A}\mathbf{x}_t\|_2^2 + \lambda\|\mathbf{x}_t\|_1\right) \qquad (3)$$

where λ is a positive regularization parameter.

Many solvers, such as l_1-magic [41], are available to solve the previous l_1-regularized optimization problem. In this paper, we employ the truncated Newton interior-point method (TNIP) proposed in [42] for its efficiency and scalability to large-scale data. It is reported that TNIP has an empirical complexity of $\mathrm{O}(n^{1.2})$, where n is the number of dictionary atoms, and that it is capable of solving million-scale problems in a few tens of minutes on a PC. Furthermore, a non-negativity constraint on the coefficients $\widehat{\mathbf{x}}_t$ can be easily added, which allows one to evaluate the importance of the non-negativity constraint for the NMF-based approaches [15]–[21].

As the sparse coefficient vector $\widehat{\mathbf{x}}_t$ becomes available, the system is ready to generate the pitch estimates for the tth frame. A note is considered active in a frame if its *activation index* is larger than a predefined threshold μ_a. Here, the activation index is defined as the sum of the absolute values of the elements in $\widehat{\mathbf{x}}_t$ corresponding to the note under consideration. Finally, the pitch of each active note is reported either in scripts or in MIDI format.

B. Construction of Dictionary

To construct the dictionary $\mathbf{A}$ for solving (3), we build a waveform database of individual notes generated by pianos with standard tuning (A4 = 440 Hz, 12-tone equal temperament). The database contains at least one waveform for each piano note and may have multiple waveforms of the same note from different pianos. In our implementation, the waveforms in the database are synthesized with three different piano timbres provided by two famous commercial software tools.[2] Then, each waveform is divided into note exemplars, and the energy of each exemplar is normalized. It is possible to divide a waveform into tens or hundreds of exemplars. However, considering the ADSR characteristics of piano sound and the computation time needed to solve (3), we only take the beginning, middle, and end parts of each waveform to form the exemplars. Since piano notes usually have a quick attack and long-lasting sustain, the beginning and middle parts of a note's waveform correspond to attack and sustain, whereas the end of the waveform corresponds to release (cf. Fig. 2). As a result, we use $D = 9$ exemplars to represent each note, giving rise to $N = 792$ exemplars in the dictionary for the 88 piano keys.

C. Pilot Study

We conducted an experiment to test the exemplar-based sparse representation method on ten one-minute long recordings of classical piano music provided by Poliner *et al.* [8]. The recordings were made using a Yamaha Disklavier piano, sampled at 8 kHz, and stored in the WAVE format [43]. We follow the Music Information Retrieval Evaluation eXchange (MIREX[3]) convention and set the frame size to 100 ms and the hop size between successive frames to 10 ms. We use magnitude spectrum as the short-time representation of the music and use the median of all activation indices as the activation threshold μ_a.

The overall estimation accuracy, evaluated in terms of the F-measure accuracy (cf. Section V-B), is much lower than that of Marolt's system (47.5% versus 66.1%) [6]. To analyze the problem, we performed a thorough examination of the results. For illustration, a sample result of the first 3 s of one test recording, Bach's *Prelude and Fugue No. 2 in C minor* is depicted in Fig. 5, where the ground truth pitches are in light gray, the pitch estimates in dark gray, and the overlaps of the ground truth and estimates in black. Three common errors were identified:

1) Chromatic error. The estimated pitch is off by one semitone. See, for example, the MIDI note 34 near frame 80 in Fig. 5. The cause of this error is that the tuning of the input music is lower than the standard tuning.
2) Octave error. The estimated pitch is off by one octave or a multiplicity of octaves. See, for example, the MIDI notes 16 and 28 near frame 25 in Fig. 5. This error occurs because notes that stand in octave relation have similar spectra.
3) Discontinuity. The estimated pitch switches between active and inactive unreasonably frequently. A pitch should normally stay active for at least 8 frames (80 ms). Any switch rate higher than the maximum rate is deemed unreasonable. See the MIDI note number 18 near frame 170 in Fig. 5. This error occurs because we treat the short-time frames independently, leaving temporal music structure unexploited.

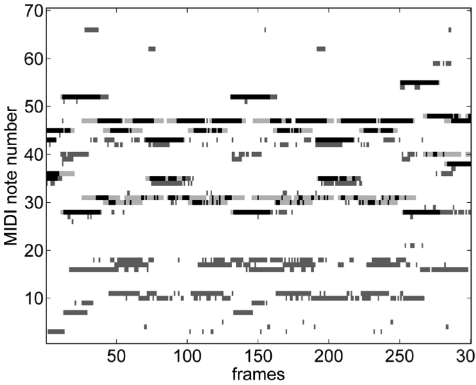


Fig. 5. Pitch estimates (dark gray) plotted on top of the ground truth (light gray). The overlaps are marked black, meaning correct pitch estimates.

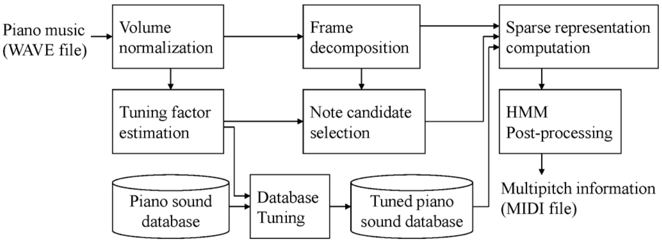


Fig. 6. Proposed multipitch estimation system.

To improve the performance, we develop three additional components which, together with the baseline scheme discussed earlier in this section, compose the final proposed system. The details of these additional components are described in the following section.

IV. Proposed System

Fig. 6 shows a schematic diagram of the proposed multipitch estimation system. The system takes a piano WAVE file as input and generates pitch estimates in MIDI file format. The system first normalizes the rms amplitude of the input signal and decomposes the normalized signal into short-time frames. To mitigate chromatic errors, a tuning factor estimation algorithm is employed and the note exemplars are tuned accordingly. Next, to reduce octave errors, we develop a note candidate selection algorithm that selects possible active notes for each frame based on the harmonic structure of piano sounds. The sparse representation of each frame is then computed based on the note candidates. Finally, we reduce the discontinuity error by exploiting the temporal relationship between frames based on HMM.

A. Tuning Factor Estimation and Dictionary Tuning

Music pieces produced by different pianos or by the same piano at different times may differ in tuning. Tuning factor estimation is needed for spectral alignment of an input music signal with note exemplars in the dictionary.

The pitch f of the ith note of a piano with a tuning factor τ is defined by

$$f = (440 \times \tau) \times \sqrt[12]{2}^{\,i-49}. \qquad (4)$$

[2] Piano #1 and Piano #2 of the Native Instrument's Bandstand (http://www.native-instruments.com) and Acoustic Piano #1 of Arobas's Guitar Pro (http://www.guitar-pro.com).

[3] MIREX is an international contest of music information retrieval tasks. http://www.music-ir.org/mirex/

TABLE I
TUNING ALGORITHM

```
Input: magnitude spectrum $\mathbf{M}$ of an exemplar, tuning factor $\tau$
Output: tuned magnitude spectrum $\mathbf{M_t}$
For $i$ = 1 to the dimension of $\mathbf{M}$
    $\mathbf{M_t}[i] = 0$
For $i$ = 1 to the dimension of $\mathbf{M}$
    $n$ = round(($i$-1) × $\tau$) + 1  // round($x$) rounds $x$ to the nearest integer
    $\mathbf{M_t}[n] = \mathbf{M_t}[n] + \mathbf{M}[i]$
```

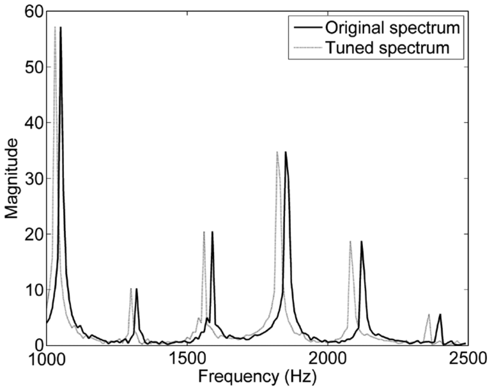


Fig. 7. Original and tuned spectra of an exemplar of C4. The estimated tuning factor is 0.982.

The value of τ is equal to 1 for standard tuning, smaller than 1 for a tuning lower than the standard tuning, and greater than 1 for a tuning higher than the standard tuning.

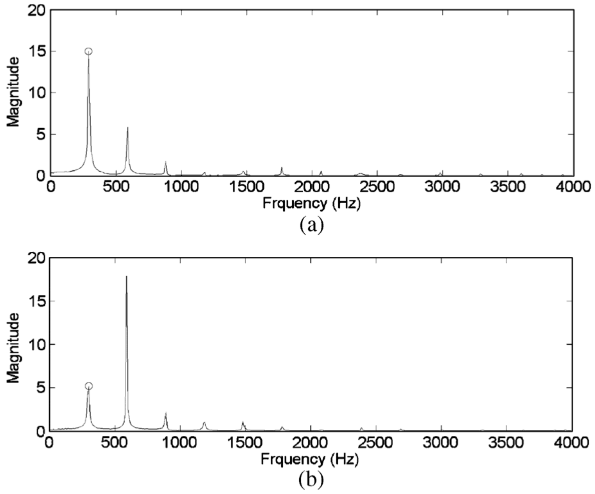
The tuning factor is estimated by the method proposed in [44] using circular statistics. After the tuning estimate is obtained, the simple tuning algorithm shown in Table I is applied to tune the note exemplars. Fig. 7 shows the result of a test example. Since the estimated tuning factor is smaller than 1, the original spectrum of a note exemplar is scaled to the left in proportion to the frequency as expected.

B. Note Candidate Selection

For each frame, the system selects a number of note candidates and uses the reduced dictionary (with fewer atoms) for solving (3).[4] On average, the number of selected candidates is 2.5 times the number of the polyphony level. The selection is based on the observation that the spectra of individual piano notes have the maximum magnitude at one of its lower harmonics, mostly the first or the second. Excluding the note exemplars corresponding to the higher harmonics from the candidate set helps reduce the octave errors.

However, we cannot simply use the spectral peak corresponding to the lowest frequency as the candidate note because the spectral magnitude of a note at the fundamental frequency may be smaller than those at the harmonic frequencies (see Fig. 8). These notes are said to be of the weak fundamental type as opposed to the strong fundamental type.

To address this problem, we develop an algorithm that first detects significant spectral peaks in a manner similar to the one proposed in [32] (see Fig. 9) and then identifies strong fundamentals and weak fundamentals from the significant spectral peaks. A spectral peak is considered significant if both of the following two criteria are satisfied: 1) its magnitude is larger than a predefined *global* threshold, μ_g and 2) its magnitude is *locally* higher than the magnitude of its smoothed version by a predefined margin μ_l. The smoothed spectrum is obtained by convolving the original spectrum with a moving Gaussian filter of radius r.



(a)

(b)

Fig. 8. Magnitude spectra of note D4 (MIDI note number 62) of (a) strong-fundamental type and (b) weak-fundamental type generated by two different pianos. The magnitude of the spectrum at the fundamental frequency is marked by a circle.

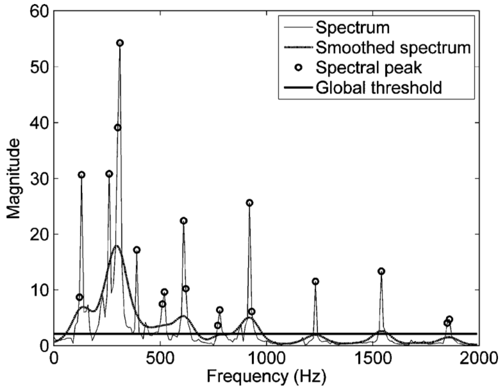


Fig. 9. Illustration of peak detection.

It is easier to identify the strong fundamentals first. Specifically, if the magnitude of the significant spectral peak at frequency f_p is higher than the magnitudes of the spectrum at all integer multiples (harmonics) of f_p, the note whose pitch is closest to f_p is selected as a candidate of the strong-fundamental type.

Then, we select candidates of the weak-fundamental type. If the magnitude of the spectrum at f_p/q, where q is an integer and is higher than a predefined threshold μ_w, the note whose pitch is closest to f_p/q is selected. In our implementation, we find it sufficient to consider two values, 2 and 3, for q.

[4] In Western music, notes are the basic building blocks and their pitches are fixed, so finding note candidates is analogous to finding pitch candidates.

C. Temporal Smoothing Based on the Hidden Markov Model

Intuitively, if a note is found active in a certain frame, it is very likely that the note is also active in the subsequent frames because a note typically lasts at least 80 milliseconds. In addition, according to the harmony theory of music, notes in harmonic relation are more likely to appear consecutively in a music piece. For example, a transition from C4 to G5 is much more likely than a transition from C4 to G#5. We refer to these two temporal structures as *intra-note smoothness and inter-note smoothness*, respectively. Both of them can be modeled by the HMM.

Two variants of HMM are considered in this paper. The first one models the intranote smoothness of each note independently using a two-state (on-and-off) HMM [8]. For the ith note, the problem can be formulated as

$$\widehat{S}^i = \arg\max_{S^i} \prod_{t=1}^{T} p\left(o_t | s_t^i\right) p\left(s_t^i | s_{t-1}^i\right) \tag{5}$$

where S^i is a state sequence, s_t^i is the state of the note at time t, o_t is the music frame beginning at time t, $p(o_t|s_t^i)$ is the probability of observing o_t given s_t^i, and $p(s_t^i|s_{t-1}^i)$ is the transition probability between states. Although we do not know $p(o_t|s_t^i)$, from the conditional probability, we have

$$p\left(s_t^i | o_t\right) \propto p\left(o_t | s_t^i\right) p\left(s_t^i\right). \tag{6}$$

Therefore, we solve

$$\widehat{S}^i = \arg\max_{S^i} \prod_{t=1}^{T} \frac{p\left(s_t^i | o_t\right)}{p\left(s_t^i\right)} p\left(s_t^i | s_{t-1}^i\right) \tag{7}$$

instead of (5). The value of $p(s_t^i|o_t)$ is obtained by dividing the activation index of the note by the maximum activation index at time t. Both the prior $p(s_t^i)$ and the state transition probability $p(s_t^i|s_{t-1}^i)$ can be learned from the training data in MIDI format. Note that the learning process is only about the intranote smoothness in this problem formulation.

The second approach models both intranote and internote smoothness by coupling HMMs, which is computationally more expensive. For K notes, we solve

$$\widehat{S}^{\{K\}} = \arg\max_{S^{\{K\}}} \prod_{i=1}^{K} \left(\prod_{t=1}^{T} p\left(o_t | s_t^i\right) \prod_{j=1}^{K} p\left(s_t^i | s_{t-1}^j\right) \right) \tag{8}$$

where $S^{\{K\}}$ is a set of K state sequences, $p(o_t|s_t^i)$ is estimated as in (7), and $p(s_t^i|s_{t-1}^j)$, $i \neq j$, is the transition probability between different notes that can also be learned from the training data. Fig. 10 illustrates the coupled HMMs. Finally, the Viterbi algorithm is applied to find the state sequence that maximizes (7) or (8). The readers are referred to [45] for the implementation details.

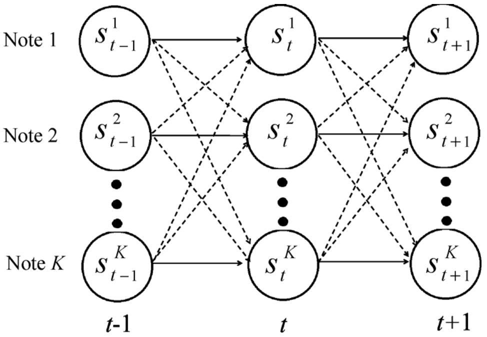


Fig. 10. Illustration of the coupled HMMs. A solid arrow indicates a transition between the states of the same note. A dashed arrow indicates a transition between the states of two notes.

V. Evaluation

We evaluate the performance of the proposed system against four state-of-the-art systems [6], [11], [13], [21]. We choose these systems for benchmarking because they offer either executable code [6], [13], [21] or test run [11], which makes performance comparison possible. Note that the systems described in [13] and [11] won the multiple fundamental frequency tracking task of MIREX in 2008 and 2010, respectively.

A. Experiment Setup

Two test data sets are used to evaluate the proposed system. The first one, referred to as Poliner10, is described in Section III-C, whereas the second one comprises the first 30 s of 60 classical piano music pieces in the MAPS database [46]. These music pieces were generated by a Steinway D piano using either ambient or close recording. We refer to the second test data set as MAPS60. Each of the 70 recordings of the two test data sets is provided with a MIDI file as the ground truth. There are a total of 4952 notes in Poliner10 and 15367 in MAPS60, with an average polyphony level of 3.5 and 3.7, respectively. The method described in Section III-B is used to construct the dictionary of note exemplars.

We use 19 1-min-long recordings described in [8] as a development data set and perform a grid search to determine the parameter values: 20 for r, 1 and 0.75 for μ_g and μ_l, respectively, 1 for μ_w, and 100 for λ. We normalize the atoms $\mathbf{a}$ using the unit l_2-norm. Note that the Poliner10 data set and the development data set are disjointed, although the recording conditions are the same. In addition, the prior and the transition probabilities in (7) and (8) are learned from a commercially available dataset,[5] consisting of 1000 MIDI transcripts of classical music, 4110 min in total.

B. Frame-Based Evaluation

Each recording is divided into 100-ms frames with a hop size of 10 ms. Three standard metrics, namely precision (P), recall (R), and F-measure (F) are used for the frame-based evaluation. These metrics are defined as follows:

$$P = \frac{N_{tp}}{N_{\mathrm{tp}} + N_{\mathrm{fp}}} \tag{9}$$

$$R = \frac{N_{\mathrm{tp}}}{N_{\mathrm{tp}} + N_{\mathrm{fn}}} \tag{10}$$

$$F = \frac{2PR}{P + R} \tag{11}$$

[5] The dataset can be ordered on http://kunstderfuge.com/.

TABLE II
FRAME-BASED EVALUATION OF THE PROPOSED SYSTEM USING BOTH DATASETS

Short-Time Representation	Precision	Recall	F-measure
Magnitude Spectrum	68.1%	71.5%	69.7%
Time Domain Signal	61.8%	71.5%	66.3%
MFCC	51.9%	69.4%	59.4%
MDCT	45.9%	69.5%	55.3%

where N_{tp} is the number of correct pitch estimates (true positives), N_{fp} is the number of inactive pitches estimated as active (false positives), and N_{fn} is the number of active pitches estimated as inactive (false negative) for all frames. It can be found that the F-measure is the harmonic mean of precision and recall; therefore, it is a better performance metric than precision and recall alone [47]. According to the common convention [10], a pitch estimate is considered correct if it is within a half semitone ($\pm 3\%$) of the ground truth.

As described in Section III-A, a non-negative constraint can be imposed on $\mathbf{x}_t$ using the TNIP-based solver. This non-negative constraint has been considered important in NMF-based methods. However, our evaluation results show that the effect of this constraint is negligible. As TNIP works slightly more efficient without the constraint, we do not confine $\mathbf{x}_t$ to be non-negative.

We first evaluate the performance of the proposed system using different short-time representations, including magnitude spectrum, time-domain signal, MFCC, and MDCT. Table II indicates that the magnitude spectrum achieves the best F-measure (69.7%) and that the time-domain signal is the second best short-time representation (F-measure 66.3%). This result may look surprising at first glance because the two representations are considered “primitive” compared to MFCC and MDCT. Yet, this result implies that sparse representation works best with a feature representation that captures all of the details of the raw signal. Due to its robustness to outliers [48], sparse representation effectively exploits the rich information contained in the magnitude spectrum or the time-domain signal for decomposition while limiting the effect of noise. As a result, we use magnitude spectrum for the short-time representation of music hereafter.

Next, we compare the performance of different multipitch estimation systems using the Poliner10 test data set. The results listed in Table III show that the proposed system achieves the best F-measure (70.2%), with an average run time comparable to Marolt’s system. Here, the run time is expressed in proportion to the real time. In addition, the coupled HMM performs slightly better than the conventional HMM, at the price of higher computational complexity. We found that the coupled HMM only brings marginal improvement because the internote smoothness has already been captured by the proposed note candidate selection algorithm. Since the selected note candidates are mostly in a harmonic relationship, using the conventional HMM for modeling intranote smoothness is already sufficient for the purpose of temporal smoothing. Therefore, the conventional HMM is adopted in the following experiments.

TABLE III
FRAME-BASED EVALUATION USING POLINER10

System	Precision	Recall	F-measure	Average Run Time
Proposed (Conventional HMMs)	74.4%	66.5%	70.2%	8.6x
Proposed (Coupled HMMs)	75.5%	67.4%	71.0%	19.1x
Marolt’s [6]	78.6%	57.1%	66.2%	8.7x
Klapuri’s [13]	72.4%	54.6%	62.3%	0.5x
Vincent’s [21]	67.2%	61.4%	64.2%	2.5x

TABLE IV
FRAME-BASED EVALUATION USING MAPS60

System	Precision	Recall	F-measure
Proposed (Default Dictionary)	66.6%	72.8%	69.6%
Proposed (Oracle Dictionary)	72.9%	72.2%	72.6%
Marolt’s	71.9%	63.4%	67.4%
Klapuri’s	66.7%	54.7%	60.1%
Vincents’s	56.3%	78.6%	65.6%

The results of our system evaluated on MAPS60 are shown in Table IV. The proposed system again achieves the best F-measure (69.6%). Moreover, when the note exemplars of the Steinway D piano are included in the dictionary (referred to as the oracle dictionary), the precision of the proposed system is improved by 6.3%, and the recall remains about the same. This important result also supports our claim that, unlike other systems, updating for a new piano can be done in our system by simply adding the note exemplars of the piano into the dictionary. In other words, our system has great adaptability.

C. Note-Based Evaluation

In addition to the frame-based evaluation, we also perform note-based performance evaluation of the proposed system against Yeh’s system [11]. The note-level pitch estimates are computed from the frame-level result as follows. If a note’s state is consecutively active for more than ten frames, the onset of the note is considered to be at the time tick 10 ms before the end of the first active frame because the hop size is 10 ms. The offset of a note is not considered because it is perceptually unimportant [8]. A note estimate is considered correct if it is off-the-ground truth by less than one frame size (i.e., 100 ms).

The evaluation is performed using Poliner10 as the test data. The number of false negatives is obtained by subtracting the number of correct note estimates from the total number of ground truth notes, and the number of false positives is obtained by subtracting the number of correct note estimates from the total number of notes output by the system. Table V shows the results of the note-based evaluation. We see that the proposed

TABLE V
NOTE-BASED EVALUATION USING POLINER10

System	Precision	Recall	F-measure
Proposed	74.6%	71.6%	73.0%
Yeh's [11]	57.2%	81.1%	67.1%

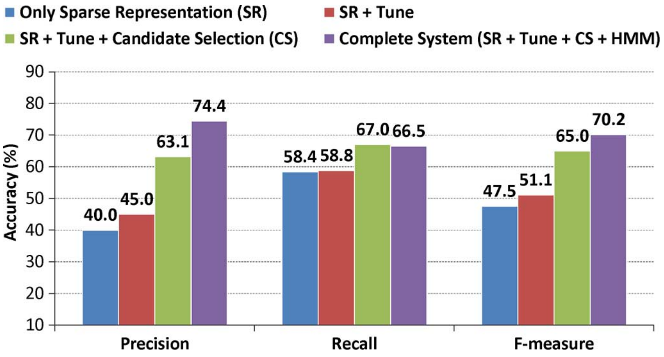


Fig. 11. Results for different system configurations.

system achieves 73.0% in F-measure, which is significantly better than Yeh's system.

In summary, the F-measure improvement of our system over the four state-of-the-art systems is significant under the one-tailed t-test ($p - \text{value} < 0.05$). Although the learning-based approaches have the upper hand in recent MIREX contests, the results of this work strongly suggest that our scheme does elevate the performance of the dictionary-based approaches to a more competitive level.

D. Analysis of System Components

To better understand the individual contribution of the system components, we evaluate the incremental performance gain by adding, one by one, the three additional components described in Section IV to the baseline system. Fig. 11 shows that the performance of the baseline system (denoted by SR) is the lowest. With tuning estimation (denoted by SR+Tune), the F-measure of the resulting system improves by 3.6%. With tuning estimation and note candidate selection (denoted by SR+Tune+Candidate Selection), the F-measure of the resulting system improves by 13.9% compared to the SR+Tune system. Finally, with all three additional components (denoted by complete system), another 5.2% gain in F-measure is obtained. We learn from this experiment that octave error is the most common error and that an effective note candidate selection significantly improves the system performance. Fig. 12 shows the result generated by the complete system for the first 3 s of Bach's Prelude and Fugue No. 2 in C minor. Compared with Fig. 5, we can clearly see that the errors, especially the octave error, are significantly reduced.

E. Number of Exemplars

To see whether the overcompleteness of the dictionary is indeed important as discussed in Section II-C, we evaluate the system performance with respect to the number of exemplars N. Three values of N are tested: 88, 264, and 792. The last value is the default setting of the proposed system. The dictionary of 264 exemplars is created by taking only the attack part of the 88 notes of the three pianos described in Section III-B, and the dictionary of 88 exemplars is created by taking the attack part of the notes of Piano #1 of Bandstand. The results shown in Fig. 13 clearly indicate that the estimation performance improves with the number of exemplars in the dictionary. Thus, we conclude that the overcompleteness of the dictionary does matter. The results also confirm our intuition that a note may not be properly represented if the number of exemplars is too small. Since the shape of the short-time spectrum of a note varies with time, we need an adequate number of exemplars to cope with the variation. Empirically, we find an average of nine exemplars per note to be good enough.

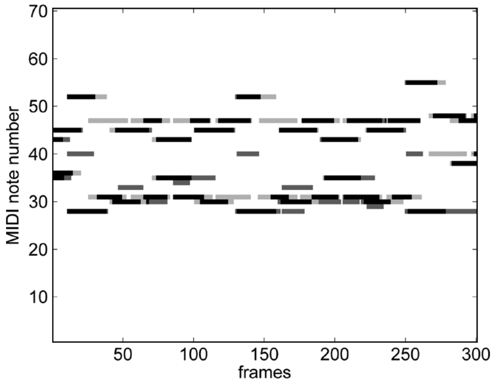


Fig. 12. Frame-level result of the first 3 s of Bach's Prelude and Fugue No. 2 in C minor.

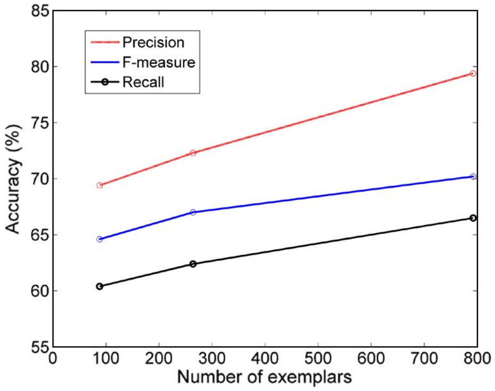


Fig. 13. Performance (in precision, F-measure, and recall, from top to bottom) of the proposed system with respect to the number of exemplars in the dictionary.

VI. CONCLUSION

We have presented a dictionary-based multipitch estimation system for piano music. It uses segments of each piano note as atoms in the dictionary. By nature, the system is free of template learning and model retraining. This makes the dictionary update an extremely easy and simple task as a new piano is considered. The novelty of the proposed system lies in its ability to manage the chromatic, octave, and discontinuity errors in an elegant and effective manner. These claims are validated through extensive evaluations. The F-measure results show that the proposed system is better than the state-of-the-art systems.

We plan to direct future work toward beat-synchronized multipitch estimation since the onsets of notes do not take place randomly but are pretty much in sync with the beats [49]. We also plan to model attack, decay, sustain, and release as separate states in the HMM to better account for the temporal variation of the note. It would also be interesting to generalize the proposed system to other instruments by taking into account that different instruments have different harmonic structures.

REFERENCES

[1] M. 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] S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press, 1990.

[3] H. F. Olson, *Music, Physics and Engineering*. New York: Dover, 1967.

[4] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, pp. 1917–1930, 2002.

[5] G. Peeters, "Music pitch representation by periodicity measures based on combined temporal and spectral representations," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, May 2006, pp. 53–56.

[6] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Trans. Multimedia*, vol. 6, no. 3, pp. 439–449, Jun. 2004.

[7] J. Bello, L. Daudet, and M. Sandler, "Automatic piano transcription using frequency and time-domain information," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2242–2251, Nov. 2006.

[8] G. Poliner and D. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP J. Adv. Signal Process.*, vol. 8, pp. 1–9, 2007.

[9] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.

[10] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 804–816, Nov. 2003.

[11] C. Yeh and A. Roebel, "Multiple-F0 estimation for MIREX 2010," Music Information Retrieval Evaluation eXchange 2010. [Online]. Available: http://www.music-ir.org/mirex/abstracts/2010/AR1.pdf

[12] C. Yeh, A. Roebel, and X. Rodet, "Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1116–1126, Aug. 2010.

[13] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Proc. Int. Conf. Music Inf. Retrieval*, Victoria, BC, Canada, Oct. 2006, pp. 216–221.

[14] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2121–2133, Nov. 2010.

[15] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, Oct. 2003, pp. 177–180.

[16] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proc. Int. Conf. Music Inf. Retrieval*, Barcelona, Spain, Oct. 2004, pp. 318–325.

[17] T. Blumensath and M. Davies, "Sparse and shift-invariant representations of music," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 50–57, Jan. 2006.

[18] A. Cont, "Realtime multiple pitch observation using sparse non-negative constraints," in *Proc. Int. Conf. Music Inf. Retrieval*, Victoria, BC, Canada, Oct. 2006, pp. 206–212.

[19] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[20] A. Dessein, A. Cont, and G. Lemaitre, "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence," in *Proc. Int. Conf. Music Inf. Retrieval*, Utrecht, The Netherlands, Aug. 2010, pp. 489–494.

[21] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

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

[23] MIDI Manufacturers Association, Complete MIDI 1.0 Detailed Specification. Nov. 2001.

[24] C. Yeh, N. Bogaards, and A. Roebel, "Synthesized polyphonic music database with verifiable ground truth for multiple-F0 estimation," in *Proc. Int. Conf. Music Inf. Retrieval*, Vienna, Austria, 2007, pp. 393–398.

[25] C. Dodge and T. A. Jerse, *Computer Music: Synthesis Composition and Performance*. New York: Schirmer, 1997, pp. 80–84.

[26] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[27] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Music genre classification via sparse representations of auditory temporal modulations," presented at the 17th Eur. Signal Process. Conf., Glasgow, Scotland, Aug. 2009.

[28] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2067–2080, Sep. 2011.

[29] C.-T. Lee, Y.-H. Yang, K.-S. Lin, and H. H. Chen, "Multiple fundamental frequency estimation of piano signals via sparse representation of Fourier Coefficients," Music Information Retrieval Evaluation eXchange. 2010. [Online]. Available: http://www.music-ir.org/mirex/abstracts/2010/LYLC1.pdf

[30] C.-T. Lee, Y.-H. Yang, and H. H. Chen, "Automatic transcription of piano music by sparse representation of magnitude spectra," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Barcelona, Spain, Jul. 2011.

[31] "Multiple F0 estimation," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, A. de Cheveigné, D. Wang, and G. J. Brown, Eds. Piscataway, NJ: IEEE/Wiley, 2006.

[32] Z. Duan, Y. Zhang, C. Zhang, and Z. Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 766–778, May 2008.

[33] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*. New York: Springer, 2006.

[34] S. A. Abdallah and M. D. Plumbley, "An independent component analysis approach to automatic music transcription," presented at the Audio Eng. Soc. 114th Convention, Amsterdam, The Netherlands, Mar. 2003.

[35] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Adv. Neural Inf. Process. Syst.*, vol. 13, pp. 556–562, 2001.

[36] S. A. Abdallah, "Towards music perception by redundancy reduction and unsupervised learning in probabilistic models," Ph.D. dissertation, Dept. Elect. Eng., King's College, London, U.K., 2002.

[37] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 17, no. 6, pp. 1142–1158, Aug. 2009.

[38] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. Davies, "Sparse representations in audio and music: From coding to source separation," *Proc. IEEE*, vol. 98, no. 6, pp. 995–1005, Jun. 2010.

[39] D. Donoho, "For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, no. 6, pp. 797–829, 2006.

[40] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.

[41] E. Candes and J. Romberg, l_1-Magic: recovery of sparse signals via convex programming, 2005. [Online]. Available: http://users.ece.gatech.edu/~justin/l1magic/

[42] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinvesky, "An interior-point method for large-scale l_1-regularized least squares," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 606–617, Dec. 2007.

[43] IBM Corporation and Microsoft Corporation, Multimedia Programming Interface and Data Spec. 1.0 Aug. 1991.

[44] K. Dressler and S. Streich, "Tuning frequency estimation using circular statistics," in *Proc. Int. Conf. Music Inf. Retrieval*, Vienna, Austria, Sep. 2007, pp. 357–360.

[45] M. Brand, "Coupled Hidden Markov Models for modeling interacting processes, mit media lab vision and modeling," Tech. Rep. 405, Nov. 1996.

[46] V. Emiya, “Transcription automatique de la musique de piano,” Ph.D. dissertation, TELECOM ParisTech, Paris, France, 2008.

[47] C. J. van Rijsbergen*, Information Retrieval*. London, U.K.: Butterworth, 1979.

[48] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, “Sparse representation for computer vision and pattern recognition,” *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010.

[49] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, “A tutorial on onset detection in music signals,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 13, no. 5, pp. 1035–1047, Sep. 2005.

Cheng-Te Lee (S’11) received the B.S. and M.S. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2008 and 2011, respectively.

His research interests include content analysis of audio signals, sound source separation, and machine learning for multimedia signal analysis.

Yi-Hsuan Yang (M’12) received the Ph.D. degree in communication engineering from National Taiwan University, Taipei, Taiwan, in 2010.

Since 2011, he has been with the Academia Sinica Research Center for Information Technology Innovation, where he is an Assistant Research Fellow. He is the author of the book *Music Emotion Recognition*. His research interests include music information retrieval, multimedia signal processing, machine learning, and affective computing.

Dr. Yang received the Microsoft Research Asia Fellowship in 2008 and the MediaTek Fellowship in 2009.

Homer H. Chen (S’83–M’86–SM’01–F’03) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana.

Since 2003, he has been with the College of Electrical Engineering and Computer Science, National Taiwan University, where he is Irving T. Ho Chair Professor. Previously, he held various R&D management and engineering positions with U.S. companies over a period of 17 years, including AT&T Bell Labs, Rockwell Science Center, iVast, and Digital Island. He was a U.S. delegate of the ISO and ITU standards committees and contributed to the development of many new interactive multimedia technologies that are now part of the MPEG-4 and JPEG-2000 standards. His research interests lie in the broad area of multimedia processing and communications.

Dr. Chen was an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology from 2004 to 2010, Associate Editor of IEEE Transactions on Image Processing from 1992 to 1994, and Guest Editor of IEEE Transactions on Circuits and Systems for Video Technology in 1999. He was also Associate Editor of *Pattern Recognition* from 1989 to 1999.