Published version: IEEE Journal of Selected Topics in Signal Processing, 5(6):1111-1123, Oct. 2011. doi: 10.1109/JSTSP.2011.2162394

Joint Multi-pitch Detection using Harmonic Envelope Estimation for Polyphonic Music Transcription

Emmanouil Benetos, *Student Member, IEEE* and Simon Dixon

***Abstract*—In this paper, a method for automatic transcription of music signals based on joint multiple-F0 estimation is proposed. As a time-frequency representation, the constant-Q resonator time-frequency image is employed, while a novel noise suppression technique based on pink noise assumption is applied in a preprocessing step. In the multiple-F0 estimation stage, the optimal tuning and inharmonicity parameters are computed and a salience function is proposed in order to select pitch candidates. For each pitch candidate combination, an overlapping partial treatment procedure is used, which is based on a novel spectral envelope estimation procedure for the log-frequency domain, in order to compute the harmonic envelope of candidate pitches. In order to select the optimal pitch combination for each time frame, a score function is proposed which combines spectral and temporal characteristics of the candidate pitches and also aims to suppress harmonic errors. For postprocessing, hidden Markov models (HMMs) and conditional random fields (CRFs) trained on MIDI data are employed, in order to boost transcription accuracy. The system was trained on isolated piano sounds from the MAPS database and was tested on classic and jazz recordings from the RWC database, as well as on recordings from a Disklavier piano. A comparison with several state-of-the-art systems is provided using a variety of error metrics, where encouraging results are indicated.**

***Index Terms*—Automatic music transcription, Harmonic envelope estimation, Conditional random fields, Resonator time-frequency image**

I. Introduction

AUTOMATIC music transcription is the process of converting an audio recording into a symbolic representation using some form of musical notation. Even for expert musicians, transcribing polyphonic pieces of music is not a trivial task, and while the problem of automatic pitch estimation for monophonic signals is considered to be a solved problem, the creation of an automated system able to transcribe polyphonic music without setting restrictions on the degree of polyphony and the instrument type still remains open. In the past years, the problem of automatic music transcription has gained considerable research interest due to the numerous applications associated with the area, such as automatic search and annotation of musical information, interactive music systems (i.e. computer participation in live human performances, score following, and rhythm tracking), as well as musicological analysis [1]–[3]. Important subtasks for automatic music transcription include pitch estimation, onset/offset detection, loudness estimation, instrument recognition, and extraction of rhythmic information. For an overview on transcription approaches, the reader is referred to [3], while in [4] a review of multiple fundamental frequency estimation systems is given.

Proposed methods for automatic transcription can be organized according to the various techniques or models employed. A large subset of the proposed systems employ signal processing techniques, usually for feature extraction, without resorting to any supervised or unsupervised learning procedures or classifiers for pitch estimation (see [3] for an overview). Several approaches for note tracking have been proposed using variants of non-negative matrix factorization (NMF), e.g. [5]. Maximum likelihood approaches, usually employing the expectation-maximization algorithm, have been also proposed in order to estimate the spectral envelope of candidate pitches or to estimate the likelihood of a set of pitch candidates (e.g. [2], [6]). Hidden Markov models (HMMs) are frequently used in a postprocessing stage for note tracking, due to the sequential structure offered by the models (e.g. [7], [8]).

Approaches for transcription related to the current work are discussed here. Yeh et al. in [9] present a multipitch estimation algorithm based on a pitch candidate set score function. The front-end of the algorithm consists of an STFT computation followed by an adaptive noise level estimation method based on the assumption that the noise amplitude follows a Rayleigh distribution. Given a pitch candidate set, the overlapping partials are detected and smoothed according to the spectral smoothness principle. The weighted score function consists of 4 features: harmonicity, mean bandwidth, spectral centroid, and synchronicity. A polyphony inference mechanism based on the score function increase selects the optimal pitch candidate set. Zhou [10] proposed an iterative method for polyphonic pitch estimation using a complex resonator filterbank as a front-end, called resonator time-frequency image (RTFI). F0 candidates are selected according to their pitch energy spectrum value and a set of rules is utilized in order to cancel extra estimated pitches. These rules are based on the number of harmonic components detected for each pitch and the spectral irregularity measure, which measures the concentrated energy around possibly overlapped partials from harmonically-related F0s.

A probabilistic method is proposed by in [6], where piano notes are jointly estimated using a likelihood function which models the spectral envelope of overtones using a smooth autoregressive (AR) model and models the residual

The authors are with the Queen Mary University of London, Centre for Digital Music, School of Electronic Engineering and Computer Science, E1 4NS London, U.K. (e-mail: emmanouilb@eecs.qmul.ac.uk; simond@eecs.qmul.ac.uk).

noise using a low-order moving average (MA) model. The likelihood function is able to handle inharmonicity and the amplitudes of overtones are assumed to be generated by a complex Gaussian random variable. In [7], Poliner and Ellis used STFT bins for frame-level piano note classification using one-versus-all support vector machines (SVMs). In order to improve transcription performance, the classification output of the SVMs was fed as input to HMMs for post-processing.

Finally, previous work by the authors includes an iterative system for multiple-F0 estimation for piano sounds [11] which incorporates temporal information for pitch estimation based on the common amplitude modulation (CAM) assumption and a public evaluation of the aforementioned system for the MIREX 2010 multiple fundamental frequency estimation task [12]. Results for the MIREX task were encouraging, considering that the system was trained on isolated piano sounds and tested on woodwind and string recordings, noting also that no note tracking procedure was incorporated.

In this work, a system for automatic transcription is proposed which is based on joint multiple-F0 estimation and subsequent note tracking. The constant-Q RTFI is used as a suitable time-frequency representation for music signals and a noise suppression method based on cepstral smoothing and pink noise assumption is proposed. For the multiple-F0 estimation step, a salience function is proposed for pitch candidate selection that incorporates tuning and inharmonicity estimation. For each possible pitch combination, an overlapping partial treatment procedure is proposed that is based on a novel method for spectral envelope estimation in the log-frequency domain, used for computing the harmonic envelope of candidate pitches. A score function which combines spectral and temporal features is proposed in order to select the optimal pitch set. Note smoothing is also applied in a postprocessing stage, employing HMMs and conditional random fields (CRFs) [13]. To the best knowledge of the authors, CRFs have not been used in the past for transcription approaches. The system was trained on a set of piano chords from the MAPS dataset [6], and tested on classic, jazz, and random piano chords from the same set, as well as on recordings from the RWC database [14], Disklavier recordings prepared in [7], and the MIREX recording used for the multiple-F0 estimation task [15]. The proposed system is compared with several approaches in the literature, where competitive results are provided using several error metrics which indicate that the current system outperforms state-of-the-art methods in many cases.

The outline of the paper is as follows. Section II describes the preprocessing steps used in the transcription system. The proposed multiple-F0 estimation method is presented in Section III. The HMM- or CRF-based postprocessing steps of the system are detailed in Section IV. In Section V, the datasets used for training and testing are presented, the employed error metrics are defined, and experimental results are shown and discussed. Finally, conclusions are drawn and future directions are indicated in Section VI, while in the Appendices a derivation for the noise suppression algorithm is given and the proposed log-frequency spectral envelope estimation method is described.

II. Preprocessing

A. Resonator Time-Frequency Image

Firstly, the input music signal is loudness-normalized to 70dB relative to the reference amplitude for 16-bit audio files, as in [16]. The resonator time-frequency image (RTFI) is employed as a time-frequency representation [10]. The RTFI selects a first-order complex resonator filter bank to implement a frequency-dependent time-frequency analysis. It can be formulated as:

$$RTFI(t,\omega) = x(t) * I_R(t,\omega) \tag{1}$$

where

$$I_R(t,\omega) = r(\omega)e^{(-r(\omega)+j\omega)t}. \tag{2}$$

$x(t)$ stands for the input signal, $I_R(t,\omega)$ is the impulse response of the first-order complex resonator filter with oscillation frequency ω and $r(\omega)$ is a decay factor which additionally sets the frequency resolution.

Here, a constant-Q RTFI is selected for the time-frequency analysis, due to its suitability for music signal processing techniques, because the inter-harmonic spacings are the same for any periodic sounds. The time interval between two successive frames is set to 40ms, which is typical for multiple-F0 estimation approaches [3]. A sampling rate of 44.1 kHz is considered for the input samples (some recordings with sampling rate 8 kHz which are presented in subsection V-A were up-converted) and the centre frequency difference between two neighboring filters is set to 10 cents (thus, the number of bins per octave b is set to 120). The frequency range is set from 27.5 Hz (A0) to 12.5 kHz (which reaches up to the 3rd harmonic of C8). The employed absolute value of the RTFI will be denoted as $X[n,k]$ from now on, where n denotes the time frame and k the log-frequency bin. When needed, $X[k]$ will stand for the RTFI slice for a single time-frame.

B. Spectral Whitening

Spectral whitening (or flattening) is a key preprocessing step applied in multiple-F0 estimation systems, in order to suppress timbral information and make the following analysis more robust to different sound sources. When viewed from an auditory perspective, it can be interpreted as the normalization of the hair cell activity level [17]. In this paper, we employ a method similar to the one in [3], but modified for log-frequency spectra instead of linear frequency ones. For each frequency bin, the power within a subband of $\frac{1}{3}$ octave span multiplied by a Hann-window $W_{hann}[k]$ is computed. The square root of the power within each subband is:

$$\sigma[k] = \left(\frac{1}{K}\sum_{l=k-K/2}^{k+K/2} W_{hann}[l]|X[l]|^2\right)^{1/2} \tag{3}$$

where $K = b/3 = 40$ bins. Afterwards, each bin is scaled according to:

$$Y[k] = (\sigma[k])^{\nu-1}X[k] \tag{4}$$

where ν is a parameter which determines the amount of spectral whitening applied and $X[k]$ is the absolute value of the RTFI for a single time frame, and $Y[k]$ is the final whitened RTFI slice. As in [3], ν was set to 0.33.

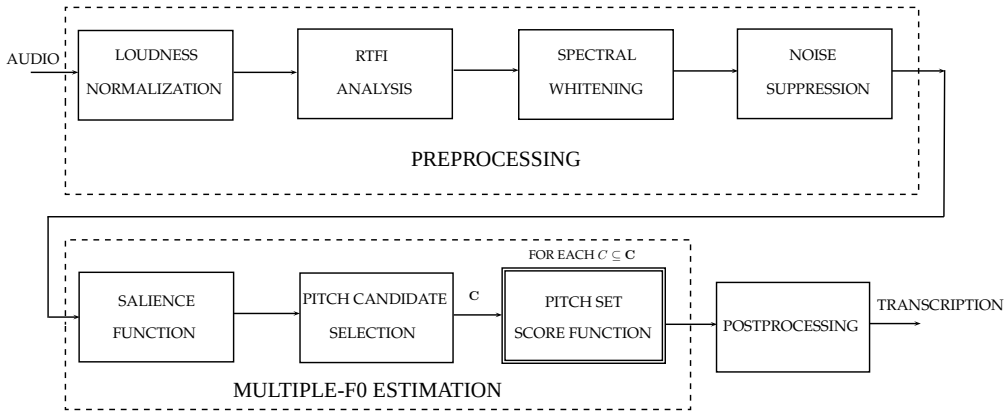

Fig. 1. Diagram for the proposed automatic transcription system.

C. Noise Suppression

In [9], an algorithm for noise level estimation was proposed, based on the assumption that noise peaks are generated from a white Gaussian process, and the resulting spectral amplitudes obey a Rayleigh distribution. Here, an approach based on pink noise assumption (elsewhere called $1/f$ noise or equal-loudness noise) is proposed. In pink noise, each octave carries an equal amount of energy, which corresponds well to the approximately logarithmic frequency scale of human auditory perception. Additionally, it occurs widely in nature, contrary to white noise and is also suitable for the employed time-frequency representation used in this work. Initial experiments were performed using a pink noise generator and the MATLAB distribution fitting toolbox. It was shown that when fitting the pink noise amplitudes with the exponential probability distribution, the resulting log likelihood was -286, compared to -539 for the Rayleigh distribution, thus motivating for the exponential distribution assumption.

The proposed signal-dependent noise estimation algorithm is as follows:

1) Perform a two-stage median filtering procedure on $Y[k]$, in a similar way to [18]. The span of the filter is set to $\frac{1}{3}$ octave. The resulting noise representation $N[k]$ gives a rough estimate of the noise level.
2) Using the noise estimate, a transformation from the log-frequency spectral coefficients to cepstral coefficients is performed [19]:
$$c_\xi = \sum_{k=1}^{K'} \log(N[k]) \cos\left(\xi\left(k-\frac{1}{2}\right)\frac{\pi}{K'}\right) \quad (5)$$
where $K' = 1043$ is the total number of log-frequency bins in the RTFI and Ξ is the number of cepstral coefficients employed, $\xi = 0, \ldots, \Xi - 1$.
3) A smooth curve in the log-magnitude, log-frequency domain is reconstructed from the first D cepstral coefficients:
$$\log|N_c(\bar{\omega})| \approx \exp\left(c_0 + 2\sum_{\xi=1}^{D-1} c_\xi \cdot \cos(\xi\bar{\omega})\right) \quad (6)$$
4) The resulting smooth curve is mapped from $\bar{\omega}$ into k. Assuming that the noise amplitude follows an exponential distribution, the expected value of the noise log amplitudes $E\{\log(|N_c(\bar{\omega})|)\}$ is equal to $\log(\lambda^{-1}) - \gamma$, where γ is the Euler constant (≈ 0.5772). Since the mean of an exponential distribution is equal to $\frac{1}{\lambda}$, the noise level in the linear amplitude scale can be described as:
$$\mathcal{L}_N(\bar{\omega}) = N_c(\bar{\omega}) \cdot e^\gamma \quad (7)$$
The analytic derivation of $E\{\log(|N_c(\bar{\omega})|)\}$ can be found in Appendix A.

In this work, the number of cepstral coefficients used was set to $D = 50$. Let $Z[k]$ stand for the whitened and noise-suppressed RTFI representation.

III. Multiple-F0 Estimation

In this section, multiple-F0 estimation, being the core of the proposed transcription system, is described. Performed on a frame-by-frame basis, a pitch salience function is generated, tuning and inharmonicity parameters are extracted, candidate pitches are selected, and for each possible pitch combination an overlapping partial treatment is performed and a score function is computed. In Fig. 1, the diagram for the proposed automatic transcription system is depicted, where the various stages for multiple-F0 estimation can be seen.

A. Salience Function Generation

In the linear frequency domain, considering a pitch p of a musical instrument sound with fundamental frequency $f_{p,1}$ and inharmonicity coefficient β_p, partials are located at frequencies:
$$f_{p,h} = h f_{p,1}\sqrt{1 + (h^2 - 1)\beta_p} \quad (8)$$
where $h \geq 1$ is the partial index [3]. Inharmonicity occurs due to string stiffness, where all partials of an inharmonic instrument have a frequency that is higher than their expected harmonic value [20]. Consequently in the log-frequency domain, considering a pitch p at bin $k_{p,0}$, overtones are located at bins:
$$k_{p,h} = k_{p,0} + \left\lfloor b \cdot \log_2(h) + \frac{b}{2}\log_2\left(1 + (h^2-1)\beta_p\right)\right\rfloor \quad (9)$$
where $b = 120$ refers to the number of bins per octave.

In addition, variations occur concerning the position of the fundamental; in [21], a model is proposed assuming that the frequency of the first partial can be shifted by a specific tuning factor. In this work, a pitch salience function $s[p, \delta_p, \beta_p]$ operating in the log-frequency domain is proposed, which incorporates tuning and inharmonicity information:

$$s[p, \delta_p, \beta_p] = \sum_{h=1}^{H} \max_{m_h} \Big\{ J[k_{p,h} + \delta_p, m_h, \beta_p] \Big\} \tag{10}$$

where

$$J[k, m_h, \beta_p] = \sqrt{Z\left[k + \left\lfloor bm_h + \frac{b}{2}\log_2(1 + (h^2 - 1)\beta_p) \right\rfloor\right]} \tag{11}$$

δ_p is the tuning deviation, and $m_h \in \mathbb{N}^*$ specifies a search range around overtone positions, belonging to the interval (m_h^l, m_h^u), where $m_h^l = \lceil \frac{\log_2(h-1)+(M-1)\log_2(h)}{M} \rfloor$, $m_h^u = \lceil \frac{(M-1)\log_2(h)+\log_2(h+1)}{M} \rfloor$. $M \in \mathbb{R}_+^*$ is a factor controlling the width of the interval, since in the log-frequency domain the search space for each harmonic is inversely proportional to the harmonic index. Here M was set to 60, so the search range for the 2nd harmonic is $[-2, 2]$ log-frequency bins, and for the 3rd and 4th harmonics is $[-1, 1]$ bins.

While the employed salience functions in the linear frequency domain (e.g. [18]) used a constant search space for each overtone, the proposed log-frequency salience function sets the search space to be inversely proportional to the partial index. The number of considered overtones H is set to 13 at maximum. The tuning deviation δ_p takes values from $[-4, \ldots, 4]$ log-frequency bins for each pitch (thus having a tuning search space of ± 40 cents around the reference tuning frequency), thus allowing the detection of notes that are not tuned using the reference frequency. The range of the inharmonicity coefficient β_p is set between 0 (completely harmonic sounds) and $5 \cdot 10^{-4}$ (moderately inharmonic sounds, e.g. from a baby grand piano [20]). The explicit modelling of inharmonicity can also be useful for temperament estimation systems, such as [22].

In order to accurately estimate the ideal tuning factor and the inharmonicity coefficient for each pitch, a 2-D maximization procedure is applied to $s[p, \delta_p, \beta_p]$ for each pitch p, in a similar manner to the work in [6]. Here $p = 1, \ldots, 88$ which corresponds to notes A0 to C8, where the pitch reference is A4 (MIDI note 69) = 440 Hz. This results in a pitch salience function estimate $s'[p]$, a tuning deviation vector and an inharmonicity coefficient vector. All in all, the computational complexity for the salience function generation is $O(N_p \cdot N_h \cdot N_\delta \cdot N_\beta)$, where $N_p = 88$, $N_h = 13$, $N_\delta = 9$, and $N_\beta = 6$ (the number of discrete values each variable takes).

Using the information extracted from the tuning and inharmonicity estimation, a harmonic partial sequence (HPS) $V[p, h]$, which contains magnitude information from $X[k]$ for each harmonic of each candidate pitch, is also stored for further processing. For example, $V[39, 2]$ corresponds to the magnitude of the 2nd harmonic of $p = 39$ (which is note B3). An example of the salience function generation is given in Fig. 2, where the RTFI spectrum of an isolated F♯3 note played by a piano is seen, along with its corresponding salience $s'[p]$. The highest peak in $s'[p]$ corresponds to $p = 34$, thus F♯3.

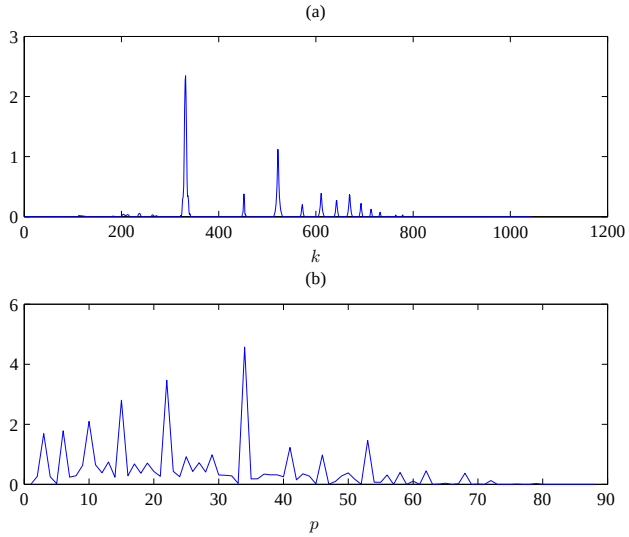

Fig. 2. (a) The RTFI slice $X[k]$ of an F♯3 piano sound. (b) The corresponding pitch salience function $s'[p]$.

B. Pitch Candidate Selection

A set of conservative rules examining the harmonic partial sequence structure of each pitch candidate is applied, which is inspired by work from [1], [23]. These rules aim to reduce the pitch candidate set for computational speed purposes. As can be seen from Fig. 2, false peaks that correspond to multiples and sub-multiples of the actual pitches occur in $s'[p]$. Here, peaks in $s'[p]$ that occur at sub-multiples of the actual F0s are subsequently deleted. In the semitone space, these peaks occur at $-\{12, 19, 24, 28, \ldots\}$ semitones from the actual pitch.

A first rule for suppressing salience function peaks is setting a minimum number for partial detection in $V[p, h]$, similar to [1]. At least three partials out of the first six need to be present in the harmonic partial sequence (since there may be a missing fundamental). A second rule discards pitch candidates with a salience value less than $0.1 \cdot \max(s'[p])$, as in [23].

Finally, after spurious peaks in $s'[p]$ have been eliminated, $C_N = 10$ candidate pitches are selected from the highest amplitudes of $s'[p]$ [6]. The set of selected pitch candidates will be denoted as $\mathbf{C}$. Thus, the maximum number of possible pitch candidate combinations that will be considered is 2^{10}, compared to 2^{88} if the aforementioned procedures were not employed. It should be stressed that this procedure does not affect the transcription performance of the system, as tested with the training set of piano chords described in subsection V-A.

C. Overlapping Partial Treatment

Current approaches in the literature rely on certain assumptions in order to recover the amplitude of overlapped harmonics. In [24], it is assumed that harmonic amplitudes decay smoothly over frequency (*spectral smoothness*). Thus, the amplitude of an overlapped harmonic can be estimated from

the amplitudes of neighboring non-overlapped harmonics. In [25], the amplitude of the overlapped harmonic is estimated through non-linear interpolation on the neighboring harmonics. In [26], each set of harmonics is filtered from the spectrum and in the case of overlapping harmonics, linear interpolation is employed.

In this work, an overlapping partial treatment procedure based on spectral envelope estimation of candidate pitches is proposed. The proposed spectral envelope estimation algorithm for the log-frequency domain is presented in Appendix B. For each possible pitch combination $C \subseteq \mathbf{C}$, overlapping partial treatment is performed, in order to accurately estimate the partial amplitudes. The proposed overlapping partial treatment procedure is as follows:

1) Given a set C of pitch candidates, estimate a partial collision list.
2) For a given harmonic partial sequence, if the number of overlapped partials is less than N_{over}, then estimate the harmonic envelope $SE_p[k]$ of the candidate pitch using only amplitude information from non-overlapped partials.
3) For a given harmonic partial sequence, if the number of overlapped partials is equal or greater than N_{over}, estimate the harmonic envelope using information from the complete harmonic partial sequence.
4) For each overlapped partial, estimate its amplitude using the harmonic envelope parameters of the corresponding pitch candidate (see Appendix B).

The output of the overlapping partial treatment procedure is the updated harmonic partial sequence $V[p,h]$ for each pitch set combination.

D. Pitch set score function

Having selected a set of possible pitch candidates and performed overlapping partial treatment on each possible combination, the goal is to select the optimal pitch combination for a specific time frame. In [9], Yeh proposed a score function which combined four criteria for each pitch: harmonicity, bandwidth, spectral centroid, and synchronicity. Also, in [23], a simple score function was proposed for pitch set selection, based on the smoothness of the pitch set. Finally, in [6] a multipitch detection function was proposed, which employed the spectral flatness of pitch candidates along with the spectral flatness of the noise residual.

Here, a weighted pitch set score function is proposed, which combines spectral and temporal characteristics of the candidate F0s, and also attempts to minimize the noise residual to avoid any missed detections. Also, features which concern harmonically-related F0s are included in the score function, in order to suppress any harmonic errors. Given a candidate pitch set $C \subseteq \mathbf{C}$ with size $|C|$, the proposed pitch set score function is:

$$\mathcal{L}(C) = \sum_{p=1}^{|C|}(\mathcal{L}_p) + \mathcal{L}_{res} \tag{12}$$

where $\mathcal{L}_p$ is the score function for each candidate pitch $p \in C$, and $\mathcal{L}_{res}$ is the score for the residual spectrum. $\mathcal{L}_p$ and $\mathcal{L}_{res}$ are defined as:

$$\begin{aligned}\mathcal{L}_p &= w_1 Fl[p] + w_2 Sm[p] - w_3 SC[p] + w_4 PR[p] - w_5 AM[p] \\ \mathcal{L}_{res} &= w_6 Fl[Res]\end{aligned} \tag{13}$$

$Fl[p]$ denotes the spectral flatness of the harmonic partial sequence:

$$Fl[p] = \frac{e^{[\sum_{h=1}^{H} \log(V[p,h])]/H}}{\frac{1}{H}\sum_{h=1}^{H} V[p,h]} \tag{14}$$

The spectral flatness is a measure of the 'whiteness' of the spectrum. Its values lie between 0 and 1 and it is maximized when the input sequence is smooth, which is the ideal case for an HPS. It has been used previously for multiple-F0 estimation in [6], [23]. Here, the definition given for the spectral flatness measure is the one adapted by the MPEG-7 framework, which can be seen in [27].

$Sm[p]$ is the *smoothness* measure of a harmonic partial sequence, which was proposed in [23]. The definition of smoothness stems from the spectral smoothness principle and its definition stems from the definition of *sharpness*:

$$Sr[p] = \sum_{h=1}^{H}(SE_p[k_{p,h}] - V[p,h]) \tag{15}$$

Here, instead of a low-pass filtered HPS using a Gaussian window as in [23], the estimated harmonic envelope SE_p of each candidate pitch is employed for the smoothness computation. $Sr[p]$ is normalized into $\bar{Sr}[p]$ and the smoothness measure $Sm[p]$ is defined as: $Sm[p] = 1 - \bar{Sr}[p]$. A high value of $Sm[p]$ indicates a smooth HPS.

$SC[p]$ is the spectral centroid for a given HPS and has been used for the score function in [9]:

$$SC[p] = \sqrt{2 \cdot \frac{\sum_{h=1}^{H} h \cdot |V[p,h]|^2}{\sum_{h=1}^{H} |V[p,h]|^2}} \tag{16}$$

It indicates the center of gravity of an HPS; for pitched percussive instruments it is positioned at lower partials. A typical value for a piano note would be 1.5 denoting that the center of gravity of its HPS is between the 1st and 2nd harmonic.

$PR[p]$ is a novel feature, which stands for the harmonically-related pitch ratio. Here, harmonically-related pitches [9] are candidate pitches in $\mathbf{C}$ that have a semitone difference of $\lceil 12 \cdot \log_2(l) \rfloor = \{12, 19, 24, 28, \ldots\}$, where $l > 1, l \in \mathbb{N}$. $PR[p]$ is applied only in cases of harmonically-related pitches, in an attempt to estimate the ratio of the energy of the smoothed partials of the higher pitch compared to the energy of the smoothed partials of the lower pitch. It is formulated as follows:

$$PR_l[p] = \sum_{h=1}^{3} \frac{V[p + \lceil 12 \cdot \log_2(l) \rfloor, h]}{V[p, l \cdot h]} \tag{17}$$

where p stands for the lower pitch and $p + \lceil 12 \cdot \log_2(l) \rfloor$ for the higher harmonically-related pitch. l stands for the harmonic relation between the two pitches ($f_{high} = l f_{low}$). In case of more than one harmonic relation between the candidate pitches, a mean value is computed: $PR[p] = \frac{1}{|N_{hr}|}\sum_{l \in N_{hr}} PR_l[p]$,

where N_{hr} is the set of harmonic relations. A high value of PR indicates the presence of a pitch in the higher harmonically-related position.

Another novel feature applied in the case of harmonically-related F0s, measuring the amplitude modulation similarity between an overlapped partial and a non-overlapped partial frequency region, is proposed. The feature is based on the common amplitude modulation (CAM) assumption, which states that partial amplitudes of a harmonic source are correlated over time [28]. Here, an extra assumption is made that frequency deviations are also correlated over time. The time-frequency region of a non-overlapped partial is compared with the time-frequency region of the fundamental. In order to compare 2-D time-frequency partial regions, the normalized tensor scalar product [29] is used:

$$AM_l[p] = \sum_{h=1}^{3} \frac{\sum_{i,j} \Lambda_{ij} B_{ij}^h}{\sqrt{\sum_{i,j} \Lambda_{ij}} \cdot \sqrt{\sum_{i,j} \Lambda_{ij} B_{ij}^h}} \quad (18)$$

where

$$\begin{aligned} \Lambda &= X[n_0 : n_1, k_{p,1} - 4 : k_{p,1} + 4] \\ B^h &= X[n_0 : n_1, k_{p,hl} - 4 : k_{p,hl} + 4] \end{aligned} \quad (19)$$

where i, j denote the indexes of matrices Λ and B^h and n_0 and $n_1 = n_0 + 5$ denote the frame boundaries of the time-frame region selected for consideration. The normalized tensor scalar product is a generalization of the cosine similarity measure, which compares two vectors, finding the cosine of the angle between them.

Res denotes the residual spectrum, which can be expressed in a similar way to the linear frequency version in [6]:

$$Res = \left\{ Z[k] \Big/ \forall p, \forall h, \left| k - k_{p,h} > \frac{\Delta_W}{2} \right| \right\} \quad (20)$$

where Δ_W denotes the mainlobe width of the employed window W. In order to find a measure of the 'whiteness' of the residual, $1 - Fl[Res]$, which denotes the residual smoothness, is used.

It should be noted that features Fl, Sr, SC, PR, AM have also been weighted by the salience function of the candidate pitch and divided by the sum of the salience function of the candidate pitch set, for normalization purposes. In order to train the weight parameters $w_i, i = 1, \ldots, 6$ of the features in (13), we used the Nelder-Mead search algorithm for parameter estimation [30]. The training set employed for experiments is described in subsection V-A. Finally, the pitch candidate set that maximizes the score function:

$$\hat{C} = \arg\max_{C \subseteq \mathbf{C}} \mathcal{L}(C) \quad (21)$$

is selected as the pitch estimate for the current frame.

IV. Postprocessing

Although temporal information has been included in the frame-based multiple-F0 estimation system, additional post-processing is needed in order to track notes over time, and eliminate any single-frame errors. In the transcription literature, hidden Markov models (HMMs) [31] have been used for postprocessing. In [32], three-state note-event HMMs were trained for each pitch, where the input features were the pitch salience value and the onset strength of the current frame. Poliner and Ellis [7] trained two-state HMMs for each note using MIDI data from the RWC database and used as observation probabilities the pseudo-posteriors of the one-versus-all SVM classifiers used for frame-based multiple-F0 estimation of piano recordings. In [33], each possible note combination between two onsets is represented by one HMM state, where the state transitions were also learned using MIDI data and the observation probability is given by the spectral flatness of the HPS of the pitch set. Finally, Cañadas-Quesada et al. also utilized two-state HMMs for each pitch that were trained using MIDI data, where the observation likelihood is given by the salience of the candidate pitch [8]. In all cases mentioned, the Viterbi algorithm is used to extract the best state sequence.

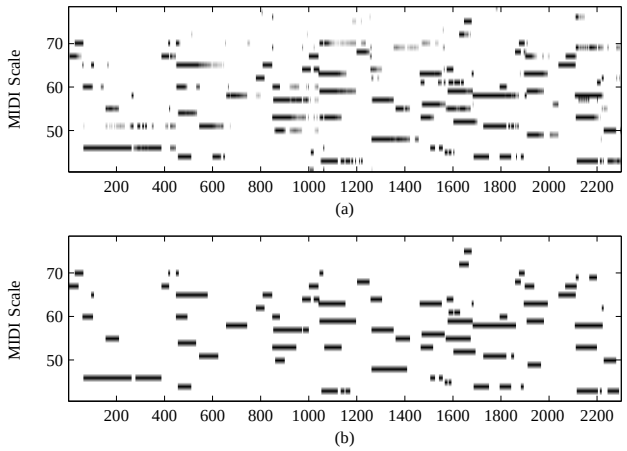

Fig. 3. Transcription output of an excerpt of 'RWC MDB-J-2001 No. 2' (jazz piano) in a 10 ms time scale (a) Output of the multiple-F0 estimation system (b) Piano-roll transcription after HMM postprocessing.

In this work, two postprocessing methods were employed: the first using HMMs and the second using conditional random fields (CRFs), which to the authors' knowledge have not been used before in music transcription research.

A. HMM Postprocessing

In this work, each pitch $p = 1, \ldots, 88$ is modeled by a two-state HMM, denoting pitch activity/inactivity, as in [7], [8]. The observation sequence is given by the output of the frame-based multiple-F0 estimation step for each pitch p: $O_p = \{o_p[n]\}, n = 1, \ldots, N$, while the state sequence is given by $Q_p = \{q_p[n]\}$. Essentially, in the HMM post-processing step, detected pitches from the multiple-F0 estimation step are tracked over time and their note activation boundaries are estimated using information from the salience function. In order to estimate the state priors $P(q_p[1])$ and the state transition matrix $P(q_p[n]|q_p[n-1])$, MIDI files from the RWC database [14] from the classic and jazz subgenres were employed, as in [8]. For each pitch, the most likely state sequence is given by:

$$Q_p' = \arg\max_{q_p[n]} \prod_n P(q_p[n]|q_p[n-1]) P(o_p[n]|q_p[n]) \quad (22)$$

in order to estimate the observation probabilities $P(o_p[n]|q_p[n])$, we employ a sigmoid curve which has as input the salience function of an active pitch from the output of the multiple-F0 estimation step:

$$P(o_p[n]|q_p[n]=1) = \frac{1}{1+e^{-(s'[p,n]-1)}} \quad (23)$$

where $s[p,n]$ denotes the salience function value at frame n. The output of the HMM-based postprocessing step is generated using the Viterbi algorithm. The transcription output of an example recording at the multiple-F0 estimation stage and after the HMM postprocessing is depicted in Fig. 3. In addition, in Fig. 4(a) the graphical structure of the employed HMMs is displayed.

B. CRF Postprocessing

Although the HMMs have repeatedly proved to be an invaluable tool for smoothing sequential data, they suffer from the limitation that the observation at a given time frame depends only on the current state. In addition, the current state depends only on its immediate predecessor. In order to alleviate these assumptions, conditional random fields (CRFs) [13] can be employed. CRFs are undirected graphical models that directly model the conditional distribution $P(Q|O)$ instead of the joint probability distribution $P(Q,O)$ as in the HMMs. This indicates that HMMs belong to the class of *generative* models, while the un-directed CRFs are *discriminative* models. The assumptions concerning the state independence and the observation dependence on the current state which are posed for the HMMs are relaxed.

In this work, 88 linear-chain CRFs are employed (one for each pitch p), where the current state $q[n]$ is dependent not only on the current observation $o[n]$, but also on $o[n-1]$. For learning, we used the same note priors and state transitions from the RWC database which were also utilized for the HMMs post-processing. For inference, the most likely state sequence for each pitch is computed using a Viterbi-like recursion which estimates:

$$Q'_p = \arg\max_{Q_p} P(Q_p|O_p) \quad (24)$$

where $P(Q_p|O_p) = \prod_n P(q_p[n]|O_p)$ and the observation probability for a given state is given as a sum of two potential functionsTS:

$$P(O_p|q_p[n]=1) = \frac{1}{1+e^{-(s'[p,n]-1)}} + \frac{1}{1+e^{-(s'[p,n-1]-1)}} \quad (25)$$

It should be noted that in our employed CRF model we assume that each note state depends only on its immediate predecessor (like in the HMMs), while the relaxed assumption over the HMMs concerns the observation potentials. The graphical structure of the linear-chain CRF which was used in our experiments is presented in Fig. 4(b).

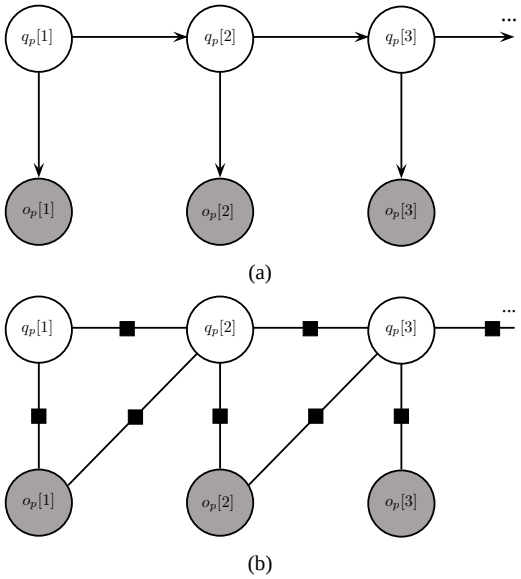

Fig. 4. Graphical structure of the employed (a) HMM (b) Linear chain CRF networks for postprocessing.

V. Evaluation

A. Datasets

For training the system parameters, samples from the MIDI Aligned Piano Sounds (MAPS) database [6] were used. The MAPS database contains real and synthesized recordings of isolated notes, musical chords, random chords, and music pieces, produced by 9 real and synthesized pianos in different recording conditions, containing around 10000 sounds in total. Recordings are stereo, sampled at 44.1 kHz, while MIDI files are provided as ground truth. Here, 103 samples from two piano types were employed for training[1], while 6832 samples from the remaining 7 piano types were used for testing on polyphonic piano sounds. The test set consists of classic, jazz, and randomly generated chords of polyphony levels 1-6, while the note range was C2-B6, in order to match the experiments performed in [6]. It should be noted that the postprocessing stage was not employed for the MAPS dataset, since it consists of isolated chords.

For the transcription experiments, we firstly used 12 excerpts from the RWC database [14], which have been used in the past to evaluate polyphonic music transcription approaches in [8], [34], [35]. A list of the employed recordings along with the instruments present in each one is shown in the top half of Table I. The recordings containing 'MDB-J' in their RWC ID belong to the jazz genre, while those that contain 'MDB-C' belong to the classic genre. For the recording titles and composer, the reader can refer to [35]. Five additional pieces were also selected from the RWC database, which have not yet been evaluated in the literature. These pieces are described in the bottom half of Table I (data 13-17). Also, the full wind quintet recording from the MIREX multi-F0 development set was also used for experiments [15]. Finally, the test dataset developed by Poliner and Ellis [7] was also used for transcription experiments. It contains 10 one-minute recordings from a Yamaha Disklavier grand piano, sampled at 8 kHz.

As far as ground-truth for the RWC data 1-12 Table I, non-aligned MIDI files are provided along with the origi-

[1]Trained weight parameters w_i were $\{1.3, 1.4, 0.6, 0.5, 0.2, 25\}$.

	RWC ID	Instruments
1	RWC-MDB-J-2001 No. 1	Piano
2	RWC-MDB-J-2001 No. 2	Piano
3	RWC-MDB-J-2001 No. 6	Guitar
4	RWC-MDB-J-2001 No. 7	Guitar
5	RWC-MDB-J-2001 No. 8	Guitar
6	RWC-MDB-J-2001 No. 9	Guitar
7	RWC-MDB-C-2001 No. 30	Piano
8	RWC-MDB-C-2001 No. 35	Piano
9	RWC-MDB-J-2001 No. 12	Flute + Piano
10	RWC-MDB-C-2001 No. 12	Flute + String Quartet
11	RWC-MDB-C-2001 No. 42	Cello + Piano
12	RWC-MDB-C-2001 No. 49	Tenor + Piano
13	RWC-MDB-C-2001 No. 13	String Quartet
14	RWC-MDB-C-2001 No. 16	Clarinet + String Quartet
15	RWC-MDB-C-2001 No. 24a	Harpsichord
16	RWC-MDB-C-2001 No. 36	Violin (polyphonic)
17	RWC-MDB-C-2001 No. 38	Violin

TABLE I
THE RWC DATA USED FOR TRANSCRIPTION EXPERIMENTS.

nal 44.1 kHz recordings. However, these MIDI files contain several note errors and omissions, as well as unrealistic note durations, thus making them unsuitable for transcription evaluation. As in [8], [34], [35], aligned ground-truth MIDI data was created for the first 23s of each recording, using Sonic Visualiser [36] for spectrogram visualization and MIDI editing. For the RWC data 13-17 in Table I, the newly-released syncRWC ground truth annotations were utilized[2].

B. Figures of Merit

In order to assess and compare the performance of the proposed system, several figures of merit from the automatic transcription literature are employed. For the piano chords using the MAPS dataset, the precision, recall, and F-measure are used:

$$Pre = \frac{tp}{tp+fp}, \quad Rec = \frac{tp}{tp+fn}, \quad F = \frac{2 \cdot Pre \cdot Rec}{Pre + Rec} \tag{26}$$

where tp is the number of correctly estimated pitches, fp is the number of false pitch detections, and fn is the number of missed pitches.

For the recordings used for the transcription experiments, several metrics are employed. It should be noted that all evaluations take place by comparing the transcribed output and the ground-truth MIDI files at a 10 ms scale, as is the standard for the multiple-F0 MIREX evaluation [15]. The first metric that is used is the overall accuracy, defined by Dixon [37]:

$$Acc_1 = \frac{tp}{fp+fn+tp} \tag{27}$$

When $Acc_1 = 1$, a perfect transcription is achieved [7]. For (27), tp, fp, and fn refer to the number of true positives, false positives, and false negatives respectively, for all frames of the recording.

A second accuracy measure is also used, which was proposed by Kameoka et al. [34] which also includes pitch substitution errors. Let $N_{ref}[n]$ stand for the number of ground-truth pitches at frame n, $N_{sys}[n]$ the number of detected pitches, and $N_{corr}[n]$ the number of correctly detected pitches. The number of false negatives at the current frame is $N_{fn}[n]$, the number of false positives is $N_{fp}[n]$, and the number of substitution errors is given by $N_{subs}[n] = \min(N_{fn}[n], N_{fp}[n])$. The accuracy measure is defined as:

$$Acc_2 = \frac{\sum_n N_{ref}[n] - N_{fn}[n] - N_{fp}[n] + N_{subs}[n]}{\sum_n N_{ref}[n]} \tag{28}$$

From the aforementioned definitions, several error metrics have been defined in [7] that measure the substitution errors (E_{subs}), miss detection errors (E_{fn}), false alarm errors (E_{fp}), and the total error (E_{tot}):

$$\begin{aligned} E_{subs} &= \frac{\sum_n \min(N_{ref}[n], N_{sys}[n]) - N_{corr}[n]}{\sum_n N_{ref}[n]} \\ E_{fn} &= \frac{\sum_n \max(0, N_{ref}[n] - N_{sys}[n])}{\sum_n N_{ref}[n]} \\ E_{fp} &= \frac{\sum_n \max(0, N_{sys}[n] - N_{ref}[n])}{\sum_n N_{ref}[n]} \\ E_{tot} &= E_{subs} + E_{fn} + E_{fp} \end{aligned} \tag{29}$$

It should be noted that the aforementioned error metrics can exceed 100% if the number of false alarms is very high [7].

C. Results

1) MAPS Database: For the isolated chord experiments using the MAPS database, the performance of the proposed transcription system compared with the results shown in [11] and [6] is shown in Fig. 5, organized according to the polyphony level of the ground truth (experiments were performed with unknown polyphony). The mean F-measures for polyphony levels $L = 1, \ldots, 6$ are 91.86%, 88.61%, 91.30%, 88.83%, 88.14%, and 69.55% respectively. It should be noted that the subset of polyphony level 6 consists only of 350 samples of random notes and not of classical and jazz chords. As far as precision is concerned, reported rates are high for all polyphony levels, ranging from 89.88% to 96.19%, with the lowest precision rate reported for $L = 1$. Recall displays the opposite performance, reaching 96.40% for one-note polyphony, and decreasing with the polyphony level, reaching 86.53%, 88.65%, 85.00%, and 83.14%, and 57.44% for levels 2-6.

In terms of a general comparison between all systems, the global F-measure for all sounds was used, where the proposed system outperforms all other approaches, reaching 88.54%. The system in [11] reports 87.47%, the system in [6] 83.70%, and finally the algorithm of [24] used for comparison in [6] reports 85.25%. By applying the same significance tests as in [11], it can be seen that the proposed method outperforms the methods of [6], [11], [24] in a statistically significant manner with 95% confidence. The aforementioned methods used for comparison follow the same pattern when Pre and Rec are concerned, reporting high Pre rates for all polyphony levels and decreasing Rec rates as polyphony increases.

[2]http://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/SyncRWC/

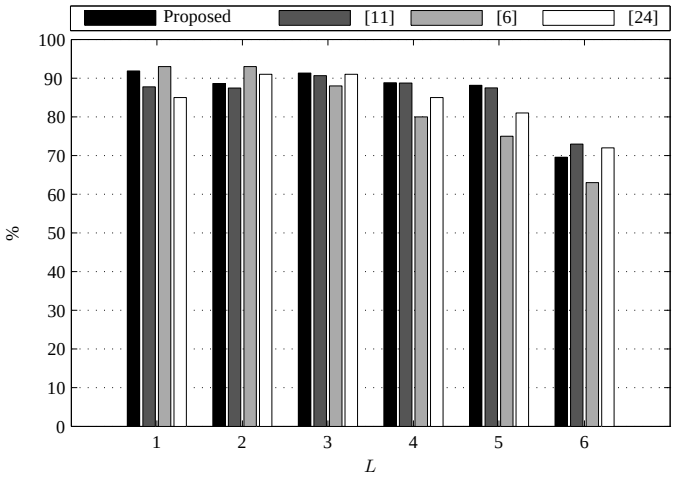

Fig. 5. Multiple-F0 estimation results for the MAPS database (in F-measure) with unknown polyphony, organized according to the ground truth polyphony level L.

2) RWC + MIREX Database: Transcription results using the RWC recordings 1-12 for the proposed system with CRF postprocessing can be found in Table II. A comparison is made using several reported results in the literature for the same files [8], [34], [35], where the proposed method reports improved mean Acc_2. Additional results were also produced for this paper using a previous method [12] submitted by the authors for the MIREX 2010 evaluation, which has a similar front-end but performs multiple-F0 estimation in an iterative fashion. Additional comparative results which demonstrate lower accuracy rates compared to the proposed system can be found in [8], that are omitted here for brevity. It should be noted that the proposed system demonstrates impressive results for some recordings compared to the state-of-the-art (e.g. in file 11, which is a cello-piano duet) while in some cases it falls behind. In file 4 for example, results are inferior compared to state-of-the-art, which could be attributed to the digital effects applied in the recording (the present system was created mostly for transcribing classical and jazz music). As far as the standard deviation of the Acc_2 metric is concerned, the proposed system reports 11.5% which is comparable to the approaches in Table II, although it is worth noting that the lowest standard deviation is reported for the method in [12].

For the RWC recordings 13-17 and the MIREX recording, transcription results can be found in Table III. It should be noted that no results have been published in the literature for these recordings. In general, it can be seen that bowed string transcriptions are more accurate than woodwind transcriptions.

Concerning the statistical significance of the proposed method's performance for the RWC recordings 1-12 compared to the various methods shown in Table II, the recognizer comparison technique described in [38] was employed. The number of pitch estimation errors of the two methods in comparison is assumed to be distributed according to the binomial law. The error rate of the proposed method is $\hat{\epsilon}_1 = E_{tot} = 0.395$, while the error rate for the methods of [8], [12], [34], [35] is $\hat{\epsilon}_2 = 0.488$, $\hat{\epsilon}_3 = 0.409$, $\hat{\epsilon}_4 = 0.438$, and $\hat{\epsilon}_5 = 0.404$, respectively. The number of examples used to generate these error rates is $\zeta = 12 \cdot 23 \cdot 100 = 27600$. Considering 95% confidence, it can be seen that $\hat{\epsilon}_i - \hat{\epsilon}_1 \geq z_{0.05}\sqrt{2\hat{\epsilon}/\zeta}$, where $i = 2, \ldots, 5$, $\hat{\epsilon} = \frac{\hat{\epsilon}_1 + \hat{\epsilon}_i}{2}$, and $z_{0.05} = 1.65$ which can be determined from tables of the Normal law. This demonstrates that the performance of the proposed transcription system is significantly better when compared with the methods in [8], [12], [34], [35]. It should be noted however that the significance threshold was only just surpassed when compared with the method of [34].

Additional insight to the proposed system's performance for all 17 RWC recordings and the MIREX one is given in Table IV, where the error metrics of subsection V-B are presented using different postprocessing configurations. It can be seen that without any postprocessing $Acc_2 = 53.8\%$, while when using the HMMs an improvement of 4.6% is reported and when the CRFs are employed, the improvement is 5.7%. It can also be seen that the note postprocessing procedures mainly decrease the number of false alarms (as can be seen in E_{fp}), at the expense however of missed detections (E_{fn}). Especially for the HMM postprocessing, a large number of missed detections have impaired the system's performance. It should be also noted that the accuracy improvement of the CRF postprocessing step over the HMM one is statistically significant with 95% confidence, using the technique in [38]. Specifically, the number of examples used to generate the error rates is $\zeta = 42200$, the error rate for the CRF postprocessing step is $\hat{\epsilon}_{CRF} = 0.405$, for the HMM step is $\hat{\epsilon}_{HMM} = 0.416$, and the significance threshold for this experiment was found to be 0.72% in terms of the error rate, which is surpassed by the CRF postprocessing (being 1.1%).

In order to test the contribution of each feature in the pitch set score function (13) to the performance of the transcription system, experiments were made on RWC recordings 1-12. For each experiment, the weight w_i, $i = 1, \ldots, 6$ in the score function that corresponds to each feature was set to 0. Results are shown in Table V, where it can clearly be seen that the most crucial feature is $Fl[Res]$, which is the residual flatness. Without that feature, the score function might select a single pitch candidate and produce several missed detections. However, it can clearly be seen that each feature significantly contributes to the final transcription result of 60.5%. When testing the contribution of the inharmonicity estimation in the salience function, the same experiment took place with no inharmonicity search, where $Acc_2 = 59.7\%$. By employing the statistical significance test of [38], the performance improvement when inharmonicity estimation is enabled is significant with 90% confidence. It should be noted however that the contribution of the inharmonicity estimation procedure depends on the instrument sources that are present in the signal. In addition, by disabling the overlapping partial treatment procedure for the same experiment, it was shown that $Acc_2 = 38.0\%$, with $E_{fp} = 20.4\%$, which indicates that false alarms from the overlapped peaks might be detected by the system. The 22.5% difference in terms of accuracy for the overlapping partial treatment is shown to be statistically significant with 95% confidence, using the method in [38].

Concerning the performance of the proposed noise suppression algorithm, comparative experiments were performed us-

	Proposed	[12]	[8]	[35]	[34]
1	60.2%	58.1%	63.5%	59.0%	64.2%
2	74.1%	50.6%	72.1%	63.9%	62.2%
3	50.0%	42.8%	58.6%	51.3%	63.8%
4	35.7%	28.8%	79.4%	68.1%	77.9%
5	75.0%	63.9%	55.6%	67.0%	75.2%
6	57.9%	52.0%	70.3%	77.5%	81.2%
7	66.8%	51.5%	49.3%	57.0%	70.9%
8	54.8%	47.0%	64.3%	63.6%	63.2%
9	74.4%	54.9%	50.6%	44.9%	43.2%
10	64.0%	58.4%	55.9%	48.9%	48.1%
11	58.9%	46.2%	51.1%	37.0%	37.6%
12	53.9%	47.6%	38.0%	35.8%	27.5%
Mean	**60.5%**	**51.2%**	**59.1%**	**56.2%**	**59.6%**
Std.	11.5%	9.0%	11.5%	12.9%	16.9%

TABLE II
TRANSCRIPTION RESULTS (Acc_2) FOR THE RWC RECORDINGS 1-12 USING THE PROPOSED METHOD WITH CRF POSTPROCESSING, COMPARED WITH OTHER APPROACHES.

	Proposed	[12]
13	48.2%	38.4%
14	41.8%	41.2%
15	66.8%	41.0%
16	70.7%	57.0%
17	75.2%	52.2%
MIREX	41.3%	39.9%
Mean	**57.4%**	**44.9%**
Std.	15.3%	7.7%

TABLE III
TRANSCRIPTION RESULTS (Acc_2) FOR RWC RECORDINGS 13-17 AND THE MIREX RECORDING, USING THE PROPOSED METHOD WITH CRF POSTPROCESSING, COMPARED WITH THE METHOD IN [12].

ing the 2-stage noise suppression procedure that was proposed for multiple-F0 estimation in [18], using the RWC recordings 1-12. The noise suppression procedure of [18] consists of median filtering on the whitened spectrum, followed by a second median filtering which does not take into account spectral peaks. Experiments with CRF postprocessing showed that transcription accuracy using the 2-state noise suppression algorithm was $Acc_2 = 56.0\%$, compared to the 60.5% of the proposed method. The performance difference is statistically significant with 95% confidence, using the method of [38].

3) Disklavier dataset [7]: Transcription results using the 10 Disklavier recording test set created by Poliner and Ellis can be found in Table VI, along with results from other approaches reported in [7]. Also, additional results were produced by the authors using our iterative MIREX-submitted method, which has a similar preprocessing front-end and the same salience function [12]. It can be seen that the best results are reported for the method in [7] while the proposed system is second-best, although it should be noted that the training set for the method by Poliner and Ellis used data from the same source as the test set. In addition, the method in [7] has displayed poor generalization performance when tested on different datasets, as can be seen from results shown in [7] and [8].

In Table VII, several error metrics are displayed for the Disklavier dataset, using different postprocessing configurations for the proposed method. The same pattern that was shown for the RWC data is shown here, where using the HMMs a small improvement of 0.4% is reported, while the improvement for the CRFs is 2.6%. The difference in the improvement over the RWC data can be attributed to the faster tempo of the Disklavier pieces. It has been argued in [8] that HMM note smoothing provides greater improvement for music pieces with slow tempo. For the HMM postprocessing, false alarms are again reduced at the expense of additional missed detections, while the CRF postprocessing displays an improvement over the missed detection errors, at the expense of false alarms.

Method	Acc_1	Acc_2	E_{tot}	E_{subs}	E_{fn}	E_{fp}
No Post.	54.4%	53.8%	46.2%	11.9%	19.4%	14.9%
HMM Post.	57.3%	58.4%	41.6%	5.4%	32.2%	4.0%
CRF Post.	58.9%	59.5%	40.5%	7.1%	25.3%	8.2%

TABLE IV
TRANSCRIPTION ERROR METRICS FOR THE PROPOSED METHOD USING RWC RECORDINGS 1-17 AND THE MIREX RECORDING, USING DIFFERENT POSTPROCESSING TECHNIQUES.

All	Fl	Sm	SC	PR	AM	$Fl[Res]$
60.5%	56.3%	59.2%	58.6%	53.5%	59.4%	29.1%

TABLE V
TRANSCRIPTION RESULTS (Acc_2) FOR THE RWC RECORDINGS 1-12 USING CRF POSTPROCESSING, WHEN FEATURES ARE REMOVED FROM THE SCORE FUNCTION (13).

VI. CONCLUSIONS

In this work, a joint multiple-F0 estimation system for automatic transcription of polyphonic music was proposed. As a front-end, the constant-Q resonator time-frequency image was selected due to its suitability for music signal representation. Contributions of the paper include:

- A noise suppression algorithm based on a pink noise assumption
- A log-frequency salience function that supports tuning and inharmonicity estimation
- Overlapping partial treatment procedure using harmonic envelopes of pitch candidates
- A pitch set score function incorporating spectral and temporal features
- An algorithm for log-frequency spectral envelope estimation based on the discrete cepstrum
- Note smoothing using conditional random fields (CRFs)

The system was trained on a set of isolated piano chords from the MAPS database and tested on recordings from the RWC database, the Disklavier database from [7], and the MIREX multipitch estimation recording [15]. Comparative results are provided using various evaluation metrics over several state-of-the-art methods, as well as on a method previously developed by the authors. The proposed system displays promising and robust results, surpassing state-of-the-art performance in many cases, considering also the fact that the training and testing datasets originate from different sources. For the RWC recordings, the improvement by the proposed system was found statistically significant compared

Method	Proposed	[11]	[7]	[32]	[39]
Acc_1	49.4%	43.3%	56.5%	41.2%	38.4%

TABLE VI
MEAN TRANSCRIPTION RESULTS (Acc_1) FOR THE RECORDINGS FROM [7] USING CRF POSTPROCESSING, COMPARED WITH OTHER APPROACHES.

Method	Acc_1	Acc_2	E_{tot}	E_{subs}	E_{fn}	E_{fp}
No Post.	46.8%	48.2%	51.8%	10.5%	35.2%	6.1%
HMM Post.	47.2%	48.3%	51.7%	8.5%	38.1%	5.1%
CRF Post.	49.4%	49.8%	50.2%	10.1%	31.4%	8.6%

TABLE VII
TRANSCRIPTION ERROR METRICS USING THE RECORDINGS FROM [7] AND DIFFERENT POSTPROCESSING TECHNIQUES.

to other approaches in the literature. For public evaluation, an iterative variant of this system was submitted for the MIREX 2010 multiple-F0 estimation task [12] displaying encouraging results, even without any postprocessing. In general, the proposed system showed improvement over the one in [12] that can be attributed to the use of pitch combinations instead of iterative selection, and the postprocessing module.

In the future, the present system will be submitted for the next MIREX evaluation. In general, results generally indicated a relatively low false alarm rate, but a considerable number of missed detections. This can be rectified in the future by relaxing several assumptions concerning the inharmonicity range and spectral smoothness (which would also allow for multipitch estimation of inharmonic instruments such as marimba or vibraphone), but at the expense of additional false positives. Also, in order to improve transcription performance, training could be applied using a multi-instrument dataset, such as the one used in [24]. In addition, more general forms of CRFs that link multiple states together could improve note prediction and smoothing. Finally, system performance can be improved by performing joint multiple-F0 estimation and note tracking, instead of frame-based multipitch estimation with subsequent note tracking.

APPENDIX A
EXPECTED VALUE OF NOISE LOG-AMPLITUDES

We assume that the noise amplitude follows an exponential distribution. In order to find the expected value of the noise log amplitudes $E\{\log(|N_c(\bar{\omega})|)\}$, we adopt a technique similar to [9]. Let $\Theta = \log(N_c(\bar{\omega})) = \Phi(N)$:

$$\begin{aligned} E\{\Theta\} &= \int_{-\infty}^{+\infty} \theta p(\theta) d\theta = \int_{-\infty}^{+\infty} \theta p(\Phi^{-1}(\theta)) \left| \frac{d\Phi^{-1}(\theta)}{d\theta} \right| \\ &= \int_{-\infty}^{+\infty} \lambda\theta e^{-\lambda e^{\theta}} e^{\theta} d\theta = \int_{0}^{+\infty} \lambda \log(\psi) e^{-\lambda\psi} d\psi \\ &= -\gamma - \lambda\log(\lambda) \cdot \int_{0}^{+\infty} e^{-\lambda\psi} d\psi \\ &= \log(\lambda^{-1}) - \gamma \end{aligned} \quad (30)$$

where γ is the Euler constant:

$$\gamma = -\int_{0}^{+\infty} e^{-\psi} \log(\psi) d\psi \approx 0.57721. \quad (31)$$

APPENDIX B
LOG-FREQUENCY SPECTRAL ENVELOPE ESTIMATION

An algorithm for posterior-warped log-frequency regularized spectral envelope estimation is proposed. Given a set of harmonic partial sequences (HPS) in the log-frequency domain, the algorithm estimates the log-frequency envelope using linear regularized discrete cepstrum estimation. In [40] a method for estimating the spectral envelope using discrete cepstrum coefficients in the Mel-scale was proposed. The superiority of discrete cepstrum over the continuous cepstrum coefficients and the linear prediction coefficients for spectral envelope estimation was argued in [41]. Other methods for envelope estimation in the linear frequency domain include a weighted maximum likelihood spectral envelope estimation technique in [42], which was employed for multiple-F0 estimation experiments in [6]. To the authors' knowledge, no other log-frequency harmonic envelope estimation algorithm has been proposed in the literature. The proposed algorithm can be outlined as follows:

1) Extract the harmonic partial sequence $V[p,h]$ and corresponding log-frequency bins $k_{p,h}$ for a given pitch p and harmonic index $h = 1, \ldots, 13$.
2) Convert the log-frequency bins $k_{p,h}$ to linear angular frequencies $\omega_{p,h}$ (where $f_s = 44.1\ kHz$ and the lowest frequency for analysis is $f_{low} = 27.5\ Hz$):
$$\omega_{p,h} = 27.5 \cdot \frac{2\pi}{f_s} \cdot 2^{\frac{k_{p,h}}{120}} \quad (32)$$
3) Perform spectral envelope estimation on $V[p,h]$ and $\omega_{p,h}$ using linear regularized discrete cepstrum (estimate coefficients $\mathbf{c}_p$). Coefficients $\mathbf{c}_p$ are estimated as:
$$\mathbf{c}_p = (\mathbf{M}_p^T \mathbf{M}_p + \varrho\mathbf{K})^{-1} \mathbf{M}_p^T \mathbf{a}_p \quad (33)$$
where $\mathbf{a}_p = [\ln(V[p,1]) \ldots \ln(V[p,H])]$, $\mathbf{K} = \text{diag}([0\ 1^2\ 2^2\ \cdots\ (K-1)^2])$, K is the cepstrum order, ϱ is the regularization parameter, and
$$\mathbf{M}_p = \begin{bmatrix} 1 & 2\cos(\omega_{p,1}) & \cdots & 2\cos(K\omega_{p,1}) \\ \vdots & \vdots & & \vdots \\ 1 & 2\cos(\omega_{p,H}) & \cdots & 2\cos(K\omega_{p,H}) \end{bmatrix} \quad (34)$$
4) Estimate the vector of log-frequency discrete cepstral coefficients $\mathbf{d}_p$ from $\mathbf{c}_p$. In order to estimate $\mathbf{d}_p$ from $\mathbf{c}_p$, we note that the function which converts linear angular frequencies into log-frequencies is given by:
$$g(\omega) = 120 \cdot \log_2\left(\frac{f_s \cdot \omega}{2\pi \cdot 27.5}\right) \quad (35)$$
which is defined for $\omega \in [\frac{2\pi \cdot 27.5}{f_s}, \pi]$. Function $g(\omega)$ is normalized using $\bar{g}(\omega) = \frac{\pi}{g(\pi)} g(\omega)$, which becomes:
$$\bar{g}(\omega) = \frac{\pi}{\log_2(\frac{f_s}{2 \cdot 27.5})} \cdot \log_2\left(\frac{f_s \cdot \omega}{2\pi \cdot 27.5}\right) \quad (36)$$
The inverse function, which converts angular log-frequencies into angular linear frequencies is given by:
$$\bar{g}^{-1}(\bar{\omega}) = \frac{2\pi \cdot 27.5}{f_s} \cdot 2^{\frac{\bar{\omega}\log_2(\frac{f_s}{2 \cdot 27.5})}{\pi}} \quad (37)$$

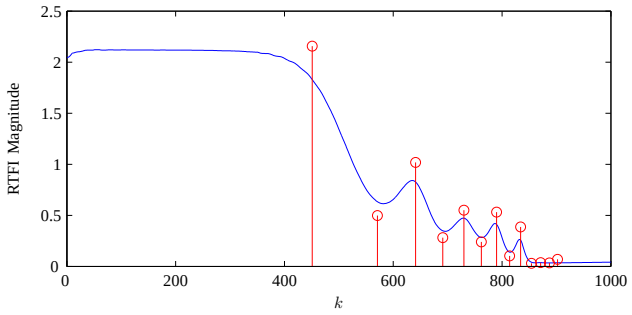

Fig. 6. Log-frequency spectral envelope of an F#4 piano tone with $P = 50$. The circle markers correspond to the detected overtones.

which is defined in $[0, \pi] \rightarrow [\frac{2\pi \cdot 27.5}{f_s}, \pi]$. From [40], it can be seen that:

$$\mathbf{d}_p = \mathbf{A} \cdot \mathbf{c}_p \tag{38}$$

where

$$\mathbf{A}_{k+1,l+1} = \frac{(2 - \delta_{0l})}{N} \sum_{n=0}^{N-1} \cos\left(l\bar{g}^{-1}(\frac{\pi n}{N})\right) \cos\left(\frac{\pi n k}{N}\right) \tag{39}$$

where N is the size of the spectrum in samples, and k, l range from 0 to $P - 1$.

5) Estimate the log-frequency spectral envelope SE from $\mathbf{d}_p$. The log-frequency spectral envelope is defined as:

$$SE_p(\bar{\omega}) = \exp\left(d_{0p} + 2 \sum_{k=1}^{P-1} d_{kp} \cos(k\bar{\omega})\right). \tag{40}$$

In Fig. 6, the warped log-frequency spectral envelope of an F#4 note produced by a piano (from the MAPS dataset) is depicted.

ACKNOWLEDGMENT

The authors would like to thank Valentin Emiya for generously providing the MAPS dataset. This work was supported by a Westfield Trust Research Studentship (Queen Mary, University of London).

REFERENCES

[1] J. P. Bello, "Towards the automated analysis of simple polyphonic music: a knowledge-based approach," Ph.D. dissertation, Department of Electronc Engineering, Queen Mary, University of London, 2003.
[2] M. Goto, "A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, pp. 311–329, 2004.
[3] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*, 2nd ed. New York: Springer-Verlag, 2006.
[4] A. de Cheveigné, "Multiple F0 estimation," in *Computational Auditory Scene Analysis, Algorithms and Applications*, D. L. Wang and G. J. Brown, Eds. IEEE Press/Wiley, 2006, pp. 45–79.
[5] P. Smaragdis, "Discovering auditory objects through non-negativity constraints," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Jeju, Korea, Oct. 2004.
[6] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.
[7] G. Poliner and D. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP J. Advances in Signal Processing*, no. 8, pp. 154–162, Jan. 2007.
[8] F. Cañadas-Quesada, N. Ruiz-Reyes, P. V. Candeas, J. J. Carabias-Orti, and S. Maldonado, "A multiple-F0 estimation approach based on Gaussian spectral modelling for polyphonic music transcription," *J. New Music Research*, vol. 39, no. 1, pp. 93–107, Apr. 2010.
[9] C. Yeh, "Multiple fundamental frequency estimation of polyphonic recordings," Ph.D. dissertation, Université Paris VI - Pierre at Marie Curie, France, Jun. 2008.
[10] R. Zhou, "Feature extraction of musical content for automatic music transcription," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Oct. 2006.
[11] E. Benetos and S. Dixon, "Multiple-F0 estimation of piano sounds exploiting spectral structure and temporal evolution," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Makuhari, Japan, Sep. 2010, pp. 13–18.
[12] ——, "Multiple fundamental frequency estimation using spectral structure and temporal evolution rules," in *Music Information Retrieval Evaluation eXchange*, Utrecht, Netherlands, Aug. 2010.
[13] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *18th Int. Conf. Machine Learning*, San Francisco, USA, Jun. 2001, pp. 282–289.
[14] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: music genre database and musical instrument sound database," in *Int. Conf. Music Information Retrieval*, Baltimore, USA, Oct. 2003.
[15] "Music Information Retrieval Evaluation eXchange (MIREX)." [Online]. Available: http://music-ir.org/mirexwiki/
[16] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Phoenix, USA, Mar. 1999, pp. 3089–3092.
[17] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, Nov. 2000.
[18] A. Klapuri, "A method for visualizing the pitch content of polyphonic music signals," in *10th Int. Society for Music Information Retrieval Conf.*, Kobe, Japan, Oct. 2009, pp. 615–620.
[19] J. C. Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," *J. Acoustical Society of America*, vol. 105, no. 3, pp. 1933–1941, Mar. 1999.
[20] L. I. Ortiz-Berenguer, F. J. Casajús-Quirós, M. Torres-Guijarro, and J. A. Beracoechea, "Piano transcription using pattern recognition: aspects on parameter extraction," in *Int. Conf. Digital Audio Effects*, Naples, Italy, Oct. 2004, pp. 212–216.
[21] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528–537, Mar. 2010.
[22] D. Tidhar, M. Mauch, and S. Dixon, "High precision frequency estimation for harpsichord tuning classification," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Dallas, USA, Mar. 2010, pp. 61–64.
[23] A. Pertusa and J. M. Iñesta, "Multiple fundamental frequency estimation using Gaussian smoothness," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Las Vegas, USA, Apr. 2008, pp. 105–108.
[24] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 6, pp. 804–816, Nov. 2003.
[25] T. Virtanen and A. Klapuri, "Separation of harmonic sounds using linear models for the overtone series," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, Orlando, USA, May 2002, pp. 1757–1760.
[26] M. R. Every and J. E. Szymanski, "Separation of synchronous pitched notes by spectral filtering of harmonics," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1845–1856, Sep. 2006.
[27] C. Uhle, "An investigation of low-level signal descriptors characterizing the noiselike nature of an audio signal," in *Audio Engineering Society 128th Convention*, London, UK, May 2010.
[28] Y. Li, J. Woodruff, and D. L. Wang, "Monaural musical sound separation based on pitch and common amplitude modulation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1361–1371, Sep. 2009.
[29] L. de Lathauwer, "Signal processing based on multilinear algebra," Ph.D. dissertation, K. U. Leuven, Belgium, 1997.
[30] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Computer J.*, vol. 7, pp. 308–313, 1965.
[31] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[32] M. Ryynänen and A. Klapuri, "Polyphonic music transciption using note event modeling," in *2005 IEEE Workshop on Applications of Signal*

Processing to Audio and Acoustics, New Paltz, USA, Oct. 2005, pp. 319–322.

[33] V. Emiya, R. Badeau, and B. David, "Automatic transcription of piano music based on HMM tracking of jointly estimated pitches," in *European Signal Processing Conf.*, Lausanne, Switzerland, Aug. 2008.

[34] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 982–994, Mar. 2007.

[35] S. Saito, H. Kameoka, K. Takahashi, T. Nishimoto, and S. Sagayama, "Specmurt analysis of polyphonic music signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 639–650, Mar. 2008.

[36] "Sonic Visualiser 1.7.1." [Online]. Available: http://www.sonicvisualiser.org/

[37] S. Dixon, "On the computer recognition of solo piano music," in *2000 Australasian Computer Music Conf.*, Jul. 2000, pp. 31–37.

[38] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik, "What size test set gives good error estimates?" *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 52–64, Jan. 1998.

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

[40] W. D'haes and X. Rodet, "Discrete cepstrum coefficients as perceptual features," in *International Computer Music Conf.*, Sep. 2003.

[41] D. Schwarz and X. Rodet, "Spectral envelope estimation and representation for sound analysis-synthesis," in *International Computer Music Conf.*, Beijing, China, Oct. 1999.

[42] R. Badeau and B. David, "Weighted maximum likelihood autoregressive and moving average spectrum modeling," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Las Vegas, USA, Apr. 2008, pp. 3761–3764.

Emmanouil Benetos (S'09) received the B.Sc. degree in informatics and the M.Sc. degree in digital media from the Aristotle University of Thessaloniki, Greece, in 2005 and 2007, respectively. In 2008, he was with the Multimedia Informatics Lab, Department of Computer Science, University of Crete, Greece. He is currently pursuing the Ph.D. degree at the Centre for Digital Music, Queen Mary University of London, U.K., in the field of automatic music transcription. His research interests include music and speech signal processing and machine learning. Mr. Benetos is a member of the Alexander S. Onassis Scholars Association.

Simon Dixon leads the Music Informatics area at the Centre for Digital Music, Queen Mary University of London. His research interests are focussed on accessing and manipulating musical content and knowledge, and involve music signal analysis, knowledge representation and semantic web technologies. He has a particular interest in high-level aspects of music such as rhythm and harmony, and has published research on beat tracking, audio alignment, chord and note transcription, characterisation of musical style, analysis of expressive performance, and the use of technology in musicology and music education. He is author of the beat tracking software BeatRoot and the audio alignment software MATCH. He was Programme Chair for ISMIR 2007, and General Co-chair of the 2011 Dagstuhl Seminar on Multimodal Music Processing, and has published over 80 papers in the area of music informatics.