POLYPHONIC MUSIC TRANSCRIPTION USING NOTE ONSET AND OFFSET DETECTION

Emmanouil Benetos and Simon Dixon

Centre for Digital Music, Queen Mary University of London, London E1 4NS, UK

ABSTRACT

In this paper, an approach for polyphonic music transcription based on joint multiple-F0 estimation and note onset/offset detection is proposed. For preprocessing, the resonator time-frequency image of the input music signal is extracted and noise suppression is performed. A pitch salience function is extracted for each frame along with tuning and inharmonicity parameters. For onset detection, late fusion is employed by combining a novel spectral flux-based feature which incorporates pitch tuning information and a novel salience function-based descriptor. For each segment defined by two onsets, an overlapping partial treatment procedure is used and a pitch set score function is proposed. A note offset detection procedure is also proposed using HMMs trained on MIDI data. The system was trained on piano chords and tested on classic and jazz recordings from the RWC database. Improved transcription results are reported compared to state-of-the-art approaches.

***Index Terms*—** Automatic transcription, multiple-F0 estimation, acoustic signal processing, music information retrieval

1. INTRODUCTION

Automatic transcription is the process of converting an audio recording into a symbolic representation using some form of musical notation. While the transcription of monophonic music 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. For an overview on transcription approaches, the reader is referred to [1].

Approaches to transcription related to the current work include the iterative spectral subtraction-based system in [1], the rule-based system in [2] which employed the resonator time-frequency image (RTFI) as a time-frequency representation, and the score function-based joint multiple-F0 estimation approach in [3]. Previous work by the authors includes a system for iterative multiple-F0 estimation [4], which was also evaluated for the 2010 MIREX multi-F0 estimation task.

As far as onset detection is concerned, an overview can be seen in [5], where the spectral flux and phase deviation are combined into a complex onset detection feature. In addition, the two aforementioned features along with an F0 descriptor are combined using decision fusion in [6].

Here, an approach for polyphonic transcription using joint multiple-F0 estimation, onset and offset detection is proposed. For onset detection, two novel descriptors are proposed which exploit information from the transcription preprocessing steps. For multiple-F0 estimation, a pitch set score function which combines several pitch-related features is proposed. Finally, novel a hidden Markov model-based offset detection procedure is proposed. Experiments on recordings from the RWC database [7] provide competitive transcription results.

The outline of the paper is as follows. In Section 2, the preprocessing steps used in the proposed system are described. The proposed onset detection procedure is presented in Section 3. Sections 4 and 5 detail the multiple-F0 estimation system and the note offset detection module, respectively. Finally, experiments are described in Section 6 and conclusions are drawn in Section 7.

This work was supported by a Westfield Trust Research Studentship (Queen Mary, University of London).

2. PREPROCESSING

2.1. Resonator Time-Frequency Image

The constant-Q resonator time-frequency image (RTFI) is employed [2], due to its suitability for music signal time-frequency representation. The RTFI selects a first-order complex resonator filter bank to implement a frequency-dependent time-frequency analysis. A constant-Q representation was selected, because the inter-harmonic spacings are the same for any periodic sounds. The time interval between two successive frames is set to 40 ms, the number of bins per octave b is set to 120, and the frequency range is set from 27.5 Hz (A0) to 12.5 kHz. From now on, the employed absolute value of the RTFI will be denoted as $X[n, k]$, where n is the time frame and k the log-frequency bin (in 10 cent resolution). When needed, $X[k]$ will stand for the RTFI slice for a single time-frame.

2.2. Spectral Whitening and Noise Suppression

Spectral whitening is applied in multiple-F0 estimation systems in order to suppress timbral information and make the following analysis more robust to different sound sources. Here, the method proposed in [1] is employed, modified for log-frequency spectra instead of linear frequency ones. For each frequency bin, the square root of the power within a subband of $\frac{1}{3}$ octave span multiplied by a Hanning window is computed, denoted $\sigma[k]$. Afterwards, each bin is scaled according to $Y[k] = (\sigma[k])^{\nu-1} X[k]$, where $\nu = 0.33$ is a parameter determining the amount of spectral whitening applied.

Afterwards, an algorithm for noise suppression is performed to the whitened RTFI. A two-stage median filtering procedure with $\frac{1}{3}$ octave span is applied to $Y[k]$ resulting in a noise representation $N[k]$, in a similar way to [4]. Cepstral smoothing using $D = 30$ coefficients is applied to $N[k]$ (as in [3]) and the resulting smooth curve $N'[k]$ is subtracted from $Y[k]$, resulting in the whitened and noise-suppressed RTFI representation $Z[k]$.

2.3. Pitch Salience Function

Using $Z[k]$, the log-frequency pitch salience function $s[p]$ proposed in [4] is extracted, where $p \in [21, \ldots, 108]$ denotes MIDI pitch. Tuning and inharmonicity coefficients are also extracted. A tuning deviation δ_p is considered for each pitch, with a tuning search space of ± 40 cents around the ideal tuning frequency. Inharmonicity is

also considered for each pitch, with the range of the inharmonicity coefficient β_p set between 0 and $5 \cdot 10^{-4}$. Using the extracted information, a harmonic partial sequence (HPS) $V[p,h]$ for each candidate pitch p and its harmonics $h = 1,\ldots,13$ is also stored for further processing.

3. ONSET DETECTION

In order to accurately detect onsets in polyphonic music, two onset descriptors which exploit information from the transcription preprocessing steps are proposed and combined using late fusion. Firstly, a novel spectral flux-based feature is defined, which incorporates pitch tuning information. Although spectral flux has been successfully used in the past for detecting hard onsets [5], false alarms may be detected for instruments that produce frequency modulations such as vibrato or portamento. Thus, a semitone-resolution filterbank is created from $Z[n,k]$, where each filter is centered at the estimated tuning position of each pitch:

$$\psi_p[p,n] = \left(\sum_{l=k_{p,0}+\delta_p-4}^{k_{p,0}+\delta_p+4} X[l,n] \cdot W_p[l] \right)^{\frac{1}{2}} \quad (1)$$

where $k_{p,0}$ is the bin that ideally corresponds to pitch p and W_p is a 80 cent-span Hanning window centered at the pitch tuning position. Using the output of the filterbank, the novel spectral flux-based descriptor is defined as:

$$SF[n] = \sum_{p=21}^{108} HW(\psi[p,n] - \psi[p,n-1]) \quad (2)$$

where $HW(\cdot) = \frac{\cdot+|\cdot|}{2}$ is a half-wave rectifier. Afterwards, onsets can be detected by performing peak picking on $SF[n]$.

In order to detect soft onsets, which may not indicate a change in signal energy [5], a pitch-based descriptor is proposed which is based on the extracted salience function. The salience function $s[p,n]$ is smoothed using a moving median filter with 120 ms span, in order to reduce any pitch fluctuations that might be attributed to amplitude modulations (e.g. tremolo). The smoothed salience function $\bar{s}[p,n]$ is then warped into a chroma-like representation:

$$Chr[\rho,n] = \sum_{i=0}^{6} \bar{s}[12 \cdot i + \rho + 20, n] \quad (3)$$

where $\rho = 1,\ldots,12$. Afterwards, the half-wave rectified first-order difference of $Chr[\rho,n]$ is used as a pitch-based onset detection function (denoted as salience difference SD):

$$SD[n] = \sum_{i=1}^{12} HW(Chr[i,n] - Chr[i,n-1]) \quad (4)$$

Accordingly, soft onsets are detected by peak picking on $SD[n]$.

In order to combine the onsets produced by the two aforementioned descriptors, late fusion is applied, as in [6]. From each of the two descriptors an onset strength signal is created, which contains either the value one at the instant of the detected onset or zero otherwise. The fused onset strength signal is created by summing and smoothing these two signals using a moving median filter of 40 ms length. Onsets are detected by performing peak picking on the fused signal by selecting peaks with a minimum 80 ms distance. For tuning onset detection parameters, a development set containing ten 30 sec classical recordings from the meter analysis data from Ghent University [8] was employed.

4. MULTIPLE-F0 ESTIMATION

4.1. Overlapping Partial Treatment

For each segment defined by two consecutive onsets, multiple-F0 estimation is applied in order to detect the pitches present. The segment is characterized by the mean $X[n,k]$ of the first 3 frames after the onset (which correspond to the steady-state part of the sound) and a corresponding segment salience function and HPS are extracted. A set of C_N candidate pitches is selected, based on the maximum values of the salience function $s[p]$ (here, C_N is set to 10 as in [9]). The pitch candidate set will be denoted as $\mathbf{C}$.

In order to recover the amplitude of overlapped harmonics, partial treatment is applied for each possible pitch candidate combination. In [1], partial amplitudes were recovered using interpolation. Here, a discrete cepstrum-based spectral envelope estimation algorithm is employed [10] in order to recover overlapped partial amplitudes. Firstly, given a set C of pitch candidates, a partial collision list is computed. For a given HPS, if the number of overlapped partials is less than N_{over}, then the amplitudes of the overlapped partials are estimated from the spectral envelope $SE_p[k]$ of the candidate pitch using only amplitude information from non-overlapped partials. If the number of overlapped partials is equal or greater than N_{over}, the partial amplitudes are estimated using spectral envelope information from the complete HPS.

4.2. 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 [3], Yeh proposed a score function which combined four criteria for each pitch: harmonicity, bandwidth, spectral centroid, and synchronicity. In addition, [9] 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:

$$\mathcal{L}(C) = \sum_{i=1}^{|C|} (\mathcal{L}_{p(i)}) + \mathcal{L}_{res} \quad (5)$$

where $\mathcal{L}_{p(i)}$ 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:

$$\begin{aligned} \mathcal{L}_p &= w_1 Fl[p] + w_2 Sm[p] - w_3 SC[p] + w_4 PR[p] \\ \mathcal{L}_{res} &= w_5 Fl[Res] \end{aligned} \quad (6)$$

$Fl[p]$ denotes the spectral flatness of the HPS, which is maximized when the input sequence is smooth and its definition can be found in [9]. $Sm[p]$ is the *smoothness* measure of a HPS, which was proposed in [11]. A high value of $Sm[p]$ indicates a smooth HPS. $SC[p]$ is the spectral centroid for a given HPS [3], which indicates its center of gravity.

$PR[p]$ is a novel feature, which stands for the harmonically-related pitch ratio. It is applied only in cases of harmonically-related

F0s in order 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:

$$PR_l[p] = \sum_{h=1}^{3} \frac{V[p + 12 \cdot \log_2(l), h]}{V[p, l \cdot h]} \quad (7)$$

where p stands for the lower pitch and $p + 12 \cdot \log_2(l)$ 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.

Res denotes the residual spectrum, which can be expressed in a similar way to the linear frequency version in [9]:

$$Res = \left\{ Z[k] \Big/ \forall p, \forall h, \left| k - k_{p,h} > \frac{\Delta_W}{2} \right| \right\} \quad (8)$$

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.

In order to train the weight parameters $w_i, i = 1, \ldots, 5$ of the features in (6), training was performed using the Nelder-Mead search algorithm for parameter estimation [12] with 100 classic, jazz, and random piano chords from the MAPS database [9] as a training set. Trained weight parameters w_i were $\{1.3, 1.4, 0.6, 0.5, 25\}$. Finally, the pitch candidate set that maximizes the score function:

$$\hat{C} = \arg\max_{C \subseteq \mathbf{C}} \mathcal{L}(C) \quad (9)$$

is selected as the pitch estimate for the current frame.

5. OFFSET DETECTION

In order to accurately detect note offsets we employ hidden Markov models (HMMs). HMMs have been used in the past for smoothing transcription results (e.g. [13]) but to the authors' knowledge they have not been utilized for offset detection. Each pitch is modeled by a two-state HMM, denoting pitch activity/inactivity. The observation sequence is given by the output of the multiple-F0 estimation step for each pitch: $O_p = \{o_p[n]\}$, $n = 1, \ldots, N$, while the state sequence is given by $Q_p = \{q_p[n]\}$. In order to estimate 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 [7] from the classic and jazz genres were used. For each pitch, the most likely state sequence is given by:

$$Q'_p = \arg\max_{q_p[n]} \prod_n P(q_p[n]|q_p[n-1]) P(o_p[n]|q_p[n]) \quad (10)$$

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:

$$P(o_p[n]|q_p[n] = 1) = \frac{1}{1 + e^{-(s[p,n]-1)}} \quad (11)$$

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 note offset is detected as the time frame when an active pitch between two consecutive onsets changes from an active to an inactive state for the first time. An example for the complete transcription system, from preprocessing to offset detection, is given in Fig. 1 for a guitar recording from the RWC database.

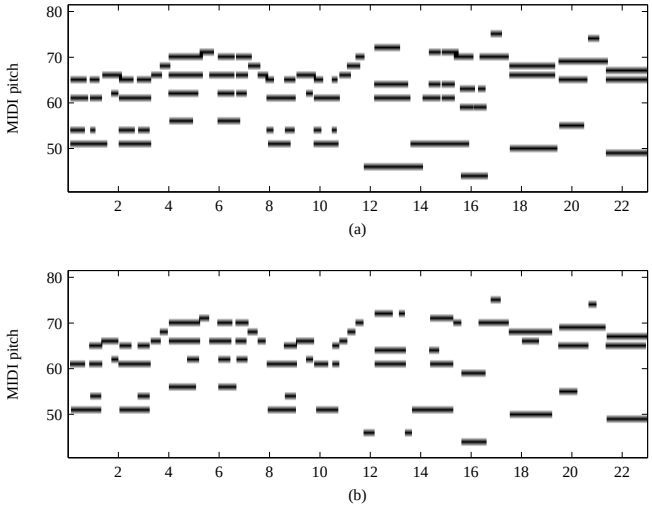

Fig. 1. (a) The pitch ground-truth of an excerpt from 'RWC MDB-J-2001 No. 9' (guitar). (b) The transcription output of the same recording. The abscissa corresponds to seconds.

6. EVALUATION

For the transcription experiments, we used 12 excerpts from the RWC database [7], which have been used in the past to evaluate transcription approaches in [14, 15, 13]. They contain classical and jazz music produced by a variety of instruments with various polyphony levels. A list of the recording titles along with the instruments present in each one can be seen in [13]. Non-aligned MIDI files are also provided as ground-truth. However, these MIDI files contain several note errors and unrealistic note durations, making them unsuitable for transcription evaluation. As in [14, 15, 13], aligned ground-truth MIDI data was created for the first 23 sec of each recording, using Sonic Visualiser (http://www.sonicvisualiser.org/). All in all, 1187 note events are contained in the test set.

For evaluating the transcription experiments, several metrics are employed, such as the overall accuracy (Acc), the total error (E_{tot}), the substitution error (E_{subs}), missed detection error (E_{fn}), and false alarm error (E_{fp}). Definitions for the aforementioned metrics can be found in [14, 15, 13]. 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. For assessing the onset detection performance of the system, the precision (Pre), recall (Rec), and F-measure (F) were employed, with a 50 ms tolerance around ground truth onset times, as in the MIREX onset detection task.

Table 1 shows transcription results for the proposed system, applying onset detection and multiple-F0 estimation only or also applying offset detection. A comparison is made with reported results in the literature for the same files [13, 15, 14], where the proposed method reports improved mean Acc. It should be noted that the proposed system demonstrates impressive results for some recordings compared to state-of-the-art (e.g. in file 10, which is string quartet recording). Additional insight to the proposed system's performance is given in Table 2, where the aforementioned error metrics are shown. It can be seen that by applying offset detection, an accuracy improvement of 1.5% is reported. Generally, the system reports relatively few false alarms, but contains a considerable number of missed detections. For comparison, excerpts from the RWC

Recording	Onsets only	Onsets+offsets	[13]	[15]	[14]
1	58.0%	60.0%	63.5%	59.0%	64.2%
2	72.1%	73.6%	72.1%	63.9%	62.2%
3	60.2%	62.5%	58.6%	51.3%	63.8%
4	64.8%	65.2%	79.4%	68.1%	77.9%
5	52.5%	53.4%	55.6%	67.0%	75.2%
6	74.4%	76.1%	70.3%	77.5%	81.2%
7	67.6%	68.5%	49.3%	57.0%	70.9%
8	58.3%	60.1%	64.3%	63.6%	63.2%
9	49.2%	50.3%	50.6%	44.9%	43.2%
10	70.5%	72.4%	55.9%	48.9%	48.1%
11	56.2%	56.2%	51.1%	37.0%	37.6%
12	33.0%	36.6%	38.0%	35.8%	27.5%
Mean	**59.7%**	**61.2%**	**59.1%**	**56.2%**	**59.6%**
Std.	11.5%	11.2%	11.5%	12.9%	16.9%

Table 1. Transcription results (Acc) for the 12 RWC recordings compared with other approaches.

Method	Acc	E_{tot}	E_{subs}	E_{fn}	E_{fp}
Onsets only	59.7%	40.3%	8.4%	24.6%	7.3%
Onsets+offsets	61.2%	38.8%	7.3%	24.8%	6.7%

Table 2. Transcription error metrics for the RWC recordings.

database are available online[1], along with synthesized transcriptions of the system.

Onset detection results using the fused descriptor, the modified SF only or the SD only, can be seen in Table 3. It should be noted that for the transcription system, we aim for high Rec instead of high F. Thus, it is more important to obtain most of the correct onsets and slightly over-segment the input (which will not affect multiple-F0 estimation), rather than lose any potential onset candidates which will lead to missed pitch detections.

7. CONCLUSIONS

In this paper, a system for automatic transcription of polyphonic music was proposed, which employed joint multiple-F0 estimation, a late fusion-based onset descriptor, and HMM-based offset detection. Experiments performed on multi-instrument recordings from the RWC database produced results which outperformed the state-of-the-art, while the use of offset detection demonstrated a consistent improvement throughout the recordings.

In the future, the proposed system will be evaluated at the forthcoming MIREX multi-F0 estimation task, as was done in 2010 for a previous system proposed by the authors [4]. In order to reduce the number of missed detections, future research will focus on modeling the attack, transient, sustain, and release states of the produced notes. Finally, system performance can be improved by performing joint multiple-F0 estimation and note tracking, instead of frame-based multiple-F0 estimation with subsequent note tracking.

8. REFERENCES

[1] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*, Springer-Verlag, New York, 2nd edition, 2006.

[1] http://www.eecs.qmul.ac.uk/~emmanouilb/transcription.html

Features	Pre	Rec	F
SF + SD	52.85%	86.84%	63.17%
SF	66.29%	81.69%	70.56%
SD	55.36%	82.42%	63.80%

Table 3. Onset detection results for the RWC recordings.

[2] R. Zhou, *Feature extraction of musical content for automatic music transcription*, Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Oct. 2006.

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

[4] 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*, Sept. 2010, pp. 13–18.

[5] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection of music signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 13, no. 5, pp. 1035–1047, Sept. 2005.

[6] A. Holzapfel and Y. Stylianou, "Three dimensions of pitched instrument onset detection," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1517–1527, 2010.

[7] 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*, Oct. 2003.

[8] M. Varewyck and J.-P. Martens, "Assessment of state-of-the-art meter analysis systems with an extended meter description model," in *8th Int. Conf. Music Information Retrieval*, 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, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.

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

[11] A. Pertusa and J. M. Iñesta, "Multiple fundamental frequency estimation using Gaussian smoothness," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2008, pp. 105–108.

[12] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Computer Journal*, vol. 7, pp. 308–313, 1965.

[13] F.J. Cañadas-Quesada, N. Ruiz-Reyes, P. Vera 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.

[14] 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.

[15] 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.