Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches

Valentin Emiya, Roland Badeau, Bertrand David

AUTOMATIC TRANSCRIPTION OF PIANO MUSIC BASED ON HMM TRACKING OF JOINTLY-ESTIMATED PITCHES

Valentin Emiya, Roland Badeau, Bertrand David

TELECOM ParisTech (ENST), CNRS LTCI
46, rue Barrault, 75634 Paris cedex 13, France
valentin.emiya@enst.fr

ABSTRACT

This work deals with the automatic transcription of piano recordings into a MIDI symbolic file. The system consists of subsequent stages of onset detection and multipitch estimation and tracking. The latter is based on a Hidden Markov Model framework, embedding a spectral maximum likelihood method for joint pitch estimation. The complexity issue of joint estimation techniques is solved by selecting subsets of simultaneously played notes within a pre-estimated set of candidates. Tests on a large database and comparisons to state-of-the-art methods show promising results.

1. INTRODUCTION

In this work, we consider the music transcription task of analyzing a recorded piece and estimating a symbolic version from it, as a MIDI file for instance. Accomplishing this task either humanly or with an automatic system often proves to be a complex issue that requires a multilevel analysis. Important subtasks include the estimation of pitch, loudness, onset and offset of notes, which can be completed by higher level analysis, such as extracting rhythm information, recognizing the played instruments or looking into the tonal context.

Many transcription systems have been developed in the recent decade, using different approaches and often mixing theoretical framework, empirical considerations and fine tuning. For instance, Davy et al. [2] and Kameoka et al. [3] use two different Bayesian approaches while the technique proposed by Marolt [4] relies on oscillators and neural networks. Blind decomposition techniques are utilized by Bertin et al. [5] and Vincent et al. [6]. The task is also accomplished by detecting temporal [1] or spectral [7] patterns with online or offline learning techniques. The system proposed by Ryynänen and Klapuri [8] takes into account the psychoacoustic knowledge of the human peripheric auditory perception and involves some musicological considerations in the postprocessing for enhancing the results.

This paper, while presenting a system whose output is a MIDI-transcribed file, focuses more on a Hidden Markov Model (HMM) framework in which a joint multiple fundamental frequency (F_0) estimation method [9] is embedded. As with most joint estimation approaches, a high number of chord combinations must be tested. Consequently, strategies are required in order to prune this search space [10]. Here, the use of an onset detector and a selector of F_0 candidates solves this computational issue by picking a reduced set of likely combinations to test.

The research leading to this paper was supported by the European Commission under contract FP6-027026-K-SPACE and by the French GIP ANR under contract ANR-06-JCJC-0027-01, *Décomposition en Éléments Sonores et Applications Musicales - DESAM*. The authors would also like to thank M. Marolt, N. Bertin, E. Vincent and M. Alonso for sharing their programs, and L. Daudet for its piano database. [1]

This paper is structured as follows. Section 2 gives an overview of the system and details each of its stages. Section 3 then reports test results and compares them with some state-of-the-art systems. Finally, conclusions are presented in section 4.

2. TRANSCRIPTION SYSTEM

The whole processing system is represented by Algorithm 1. The input is the recording of a piece of piano music, denoted by $x(t)$, the corresponding discrete sequence being sampled at $F_s = 22050$ Hz. Firstly, an onset detection is applied. The signal portion between two successive onsets will be herein referred to as a **segment**. For each of these variable-length segments, we select a presumably oversized set of F_0 candidates. The segment is then split into overlapping **frames** of constant length N, which are processed using a HMM. In this framework, the likelihood related to each possible set of simultaneous notes (or local chord) is derived using a recent spectral Maximum Likelihood (ML) estimation technique [9]. A MIDI file is generated at the end.

Algorithm 1 (System overview)

Input: Waveform
  Detect onsets
  for each segment between subsequent onsets **do**
    Select note candidates
    Track the most likely combination of notes within the HMM framework
  end for
  Detect repetitions of notes from one segment to the next one
Output: MIDI file

2.1 Onset detection

Onset detection is performed by an existing algorithm [11] based on the so-called spectral energy flux. An adaptive threshold is used to extract the local maxima of a detection function that measures phenomenal accents. The temporal accuracy of the onset detection is about 12 ms.

2.2 Selection of F_0 candidates

In order to select the set of note candidates, a normalized product spectrum is derived as a function of the fundamental frequency. The function is defined by:

$$S(f_0) = \frac{1}{H(f_0)^{\nu}} \ln \prod_{h=1}^{H(f_0)} |X(f_h)|^2 \qquad (1)$$

where $H(f_0)$ is the number of overtones below a predefined cut-off frequency for fundamental frequency f_0, $X(f)$ is a whitened version of the Discrete Fourier Transform (DFT) of the current frame, ν is a parameter empirically

set to .38 to adjust the normalization term $H(f_0)^\nu$ and f_h is the frequency of the overtone with order h defined by $f_h = hf_0\sqrt{1+\beta h^2}$, β being the so-called inharmonicity coefficient of the piano tone [12]. The whitening process reduces the spectral dynamics. It is performed by modeling the background noise with an autoregressive process and by applying the inverse filter to the original signal frame. The normalization by $H(f_0)^\nu$ aims at correcting the slope due to the variation of the number of overtones with respect to f_0. In the special case $\nu = 0$, the function in (1) equals the product spectrum [13], which has been designed for a constant number of overtones. For each note in the search range, the function is maximized in a 2-dimensional plane in the neighborhood of $(f_0, \beta(f_0))$ where f_0 is the well-tempered scale value of the fundamental frequency and $\beta(f_0)$ is a typical value of the inharmonicity coefficient for this note [12, pp. 365]. The N_c notes with the highest values are then selected as candidates. In practice, the function is computed and averaged over the first three frames of each segment in order to be robust to the temporal spreading of onsets before the steady state of notes is reached. N_c is set to 9, the frame length is 93 ms, with a 50%-overlap. The cut-off frequency related to $H(f_0)$ varies from 1200 Hz in the bass range ($f_0 = 65$ Hz) to $F_s/2$ in the treble range ($f_0 = 1000$ Hz and above).

The candidate selection stage has two purposes. Firstly, as with all joint estimation techniques, the approach described in this paper faces a high combinatory issue. Selecting a reduced set of candidates is a way to reduce the number of combinations to be tested, and to reach a realistic computational cost. For instance, if the maximum polyphony is 5 and the note search range spreads over 5 octaves (*i.e.* 60 notes), around $5 \cdot 10^6$ combinations have to be tested, whereas after selecting $N_c = 9$ candidates among the 60 notes, the size of the set of possible chords decreases to only 382 elements. Secondly, the efficiency of the transcription system depends on the selection of optimized values for fundamental frequencies and inharmonicity. These values can be obtained by either maximizing the candidate selection function defined above, which is a criterion on the energy of overtones, or maximizing the likelihood described in the next section. The advantage of the first solution is that it reduces the overall computational cost: the optimization task is performed on a lower number of candidates (*e.g.* with the figures mentioned above, 60 note candidates instead of 382 chords) and a call to the candidate selection function does not require as much computation time as a call to the likelihood function.

2.3 HMM tracking of most likely notes

This section describes how to track the possible combinations of note candidates, along frames, by means of one HMM [14] per segment. From now on, each possible combination of notes in a frame is referred to as a *chord*.

Let us consider a segment obtained by the onset detection stage, *i.e.* delimited by two consecutive onsets. It is composed of U frames, numbered from 1 to U. In frame u, $1 \leq u \leq U$, the observed spectrum X_u is a random variable that depends on the underlying chord, denoted by c_u. c_u is also a random variable. When one or several notes are played, they spread over a number of consecutive frames and may be extinguished at any moment. Thus, in a short-term context like the duration of a segment, the polyphonic content c_u of frame u strongly depends on the short-term past. Consequently, a first-order Markovian process is assumed for the chord sequence $c_1 \dots c_U$: for $u \geq 2$, chord c_u only depends on chord c_{u-1} and does not depend on u, resulting in

$$p(c_{u+1}|c_1 \dots c_u) = p(c_{u+1}|c_u) \tag{2}$$

Thus the transcription of the segment consists in finding the best sequence of hidden chords $\hat{c}_1\hat{c}_2\dots\hat{c}_U$ given the sequence of observations $X_1X_2\dots X_U$.

In statistical terms, $\hat{c}_1 \dots \hat{c}_U$ is obtained by solving:

$$\hat{c}_1 \dots \hat{c}_U = \underset{c_1 \dots c_U}{\operatorname{argmax}}\, p(c_1 \dots c_U|X_1 \dots X_U) \tag{3}$$

which is rewritten using the Bayes rule as:

$$\hat{c}_1 \dots \hat{c}_U = \underset{c_1 \dots c_U}{\operatorname{argmax}}\, \frac{p(c_1 \dots c_U)\, p(X_1 \dots X_U|c_1 \dots c_U)}{p(X_1 \dots X_U)} \tag{4}$$

The denominator is removed without changing the argmax result:

$$\hat{c}_1 \dots \hat{c}_U = \underset{c_1 \dots c_U}{\operatorname{argmax}}\, p(c_1 \dots c_U)\, p(X_1 \dots X_U|c_1 \dots c_U) \tag{5}$$

Given that the observed spectrum X_u only depends on the underlying chord c_u, we obtain:

$$\hat{c}_1 \dots \hat{c}_U = \underset{c_1 \dots c_U}{\operatorname{argmax}}\, p(c_1 \dots c_U) \prod_{u=1}^{U} p(X_u|c_u) \tag{6}$$

Using eq. (2), this finally reduces to:

$$\hat{c}_1 \dots \hat{c}_U = \underset{c_1 \dots c_U}{\operatorname{argmax}}\, p(c_1) \prod_{u=2}^{U} p(c_u|c_{u-1}) \prod_{u=1}^{U} p(X_u|c_u) \tag{7}$$

In a HMM context, c_u is the so-called hidden state in frame u, X_u is the observation, $p(c_1)$ is the initial-state probability, $\lambda_{c_{u-1},c_u} \triangleq p(c_u|c_{u-1})$ is the state-transition probability from c_{u-1} to c_u and $p(X_u|c_u)$ is the observation probability in state c_u. Thanks to the selection of N_c note candidates in the current segment and to the polyphony limit set to $P_{\max}$, possible values for c_u are restricted to the sets composed of 0 to $P_{\max}$ notes among the candidates. Thus, the number of states equals $\sum_{P=0}^{P_{\max}} \binom{N_c}{P}$. Transition probabilities are defined as follows:

- for $2 \leq u \leq U$, the birth of a note is not allowed in frame u since a segment is defined as being delimited by two consecutive onsets. Thus $\lambda_{c,c'} \triangleq 0$ if c' is not a subset of notes from c.
- as a result, the only transition from the "silence" state is toward itself: $\lambda_{\emptyset,\emptyset} \triangleq 1$.
- transitions from c are allowed toward c itself or toward a subset c' of c. It only depends on the number of notes in chord c and in chord c'. Probability transitions are learnt from musical pieces, as described below.

The initial-state probabilities are also learnt as a function of the number of notes in the state. The learning process is performed on MIDI music files. The initial-state probabilities are learnt from the polyphony observed at each onset time and the transition probabilities from the polyphony evolution between consecutive onset times. Finally, $p(X_u|c_u)$ is the likelihood detailed in the next section and given in a logarithmic version in eq. (13): $\ln p(X_u|c_u) = \tilde{L}_{X_u}(c_u)$.

The Viterbi algorithm [15] is applied to extract the best sequence of states that explains the observations, as illustrated in Table 1 and Figure 1. The obtained chord sequence corresponds to the set of notes present at the beginning of the segment and to the possible terminations within the segment.

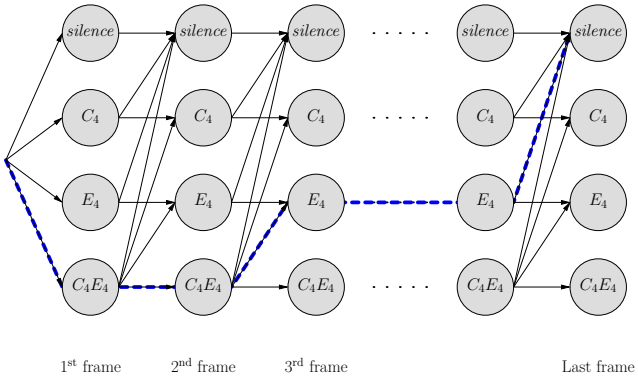

Figure 1: Chord network corresponding to the transition matrix of Table 1. Due to the sparsity of the transition matrix, transitions are allowed toward the same chord or toward a "subchord", when note endings occur. The thick, dashed line shows a possible Viberbi path: chord $\{C_4, E_4\}$ is detected, C_4 dies at frame 3 while E_4 lasts until next to last frame of the segment.

c \ c'	$\emptyset$	C_4	E_4	C_4 E_4
$\emptyset$	1	0	0	0
C_4	.17	.83	0	0
E_4	.17	0	.83	0
C_4E_4	.07	.06	.06	.80

Table 1: Example of transition matrix, *i.e.* the probability of going from chord c at time t to chord c' at time $t+1$. For graphical convenience, only $N_c = 2$ candidates are selected (notes C_4 and E_4). The transition probability is learnt as a function of the number of notes in c and c'.

Note that this HMM-based approach shares some similarities with another transcription system [10] in which chords are also modeled by HMM states. However, major differences exists between the two systems, especially in the choice of the observations, of their likelihood estimation, and in the way the HMM networks are significantly simplified, thanks to the onset-based segmentation, the selection of note candidates and the use of one state per chord in the current paper.

2.4 Maximum likelihood estimation of simultaneous pitches

The core of the pitch estimation stage is performed at the frame level by estimating the likelihood of the observed DFT X given a possible chord. In 2.4.1 and 2.4.2, we present a summary of this method that has been developed in a recent work [9]. A normalization stage is then introduced, with a number of purposes: obtaining a homogeneous detection function with respect to low/high pitched notes, making it robust to polyphony variations, and adjusting dynamics between the note and the noise likelihoods. A silence detection mechanism is finally described.

2.4.1 Maximum likelihood principle

Let us consider a mixture of M notes and of an additive colored noise. We observe one frame from this mixture. For $1 \leq m \leq M$, note m is modeled by a set of sinusoidal components. The set of related frequency bins is denoted by $\mathcal{H}^{(m)}$. In the case of a piano note, $\mathcal{H}^{(m)}$ is defined by:

$$\mathcal{H}^{(m)} = \left\{ f_h \,\middle|\, f_h = h f_0^{(m)} \sqrt{1 + \beta^{(m)} h^2} < \frac{F_s}{2} \right\} \quad (8)$$

where $f_0^{(m)}$ is the fundamental frequency of note m and $\beta^{(m)}$ is its inharmonicity coefficient.

We then assume that the noise spectrum is observed in any frequency bin that is not located within the primary spectral lobe of a note component. The set $\mathcal{N}$ of frequency bins related to noise observations is inferred by $\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(M)}$ and is thus defined by:

$$\mathcal{N} = \left\{ f \in \mathcal{F} \,\middle|\, \forall f' \in \bigcup_{m=1}^{M} \mathcal{H}^{(m)}, |f - f'| > \Delta f/2 \right\} \quad (9)$$

where $\mathcal{F}$ is the whole set of frequency bins and Δf is the width of the primary spectral lobe ($\Delta f = 4/N$ for a Hann window).

We now consider a set $\mathcal{S} \in \left\{\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(M)}, \mathcal{N}\right\}$, *i.e.* the set of frequency bins related to any of the elements (either a note or the noise) in the mixture. We model the set of selected spectral observations $X(\mathcal{S})$ by a transfer function in a family of parametric functions (*e.g.* all-pole functions)[1]. We showed in [9] that the normalized log-likelihood of $X(\mathcal{S})$ can be analytically written as:

$$L_{\mathcal{S}}(R) = c - 1 + \ln\left(\rho_{\mathcal{S}}(R)\right) \quad (10)$$

where R is the considered parametric transfer function, $c = -\frac{1}{\#\mathcal{S}} \sum_{k \in \mathcal{S}} \ln(\pi |X(k)|^2)$ is a constant w.r.t. R ($\#\mathcal{S}$ denotes the number of elements in $\mathcal{S}$) and

$$\rho_{\mathcal{S}}(R) = \frac{\left(\prod_{k \in \mathcal{S}} \left| \frac{X(k)}{R(k)} \right|^2 \right)^{\frac{1}{\#\mathcal{S}}}}{\frac{1}{\#\mathcal{S}} \sum_{k \in \mathcal{S}} \left| \frac{X(k)}{R(k)} \right|^2} \quad (11)$$

is equal to the ratio between the geometrical mean and the arithmetical mean of the set $\left\{ \left| \frac{X(k)}{R(k)} \right|^2 \right\}_{k \in \mathcal{S}}$. Such a ratio is maximal and equal to 1 when $|X(k)/R(k)|$ is constant, independent of k, which means that $\rho_{\mathcal{S}}(R)$ measures the

[1] Note that in the case of frequency overlap between two notes, this modelisation does not hold for the overlapped frequency bins. This phenomenon is ignored here.

whiteness, or the *flatness* of $\left\{\left|\frac{X(k)}{R(k)}\right|^2\right\}_{k\in\mathcal{S}}$. Thus the application of the Maximum Likelihood principle, *i.e.* the estimation of the parameterized function that maximizes $\rho_{\mathcal{S}}(R)$ results in whitening the spectral envelope $|X(\mathcal{S})|$. In our system, the note $m \in [\![1, M]\!]$ is modeled by an autoregressive (AR) filter whereas noise is modeled as a finite impulse response (FIR) filter of length $p \ll N$. A discussion on the choice of these models is presented in [9]. Approximate solutions $\widehat{R}_{\mathcal{H}_1}, \ldots, \widehat{R}_{\mathcal{H}_M}, \widehat{R}_{\mathcal{N}}$ to the optimization problem are obtained by means of estimation techniques in the case of partially observed spectra [16].

2.4.2 Joint estimation function

The M simultaneous notes are parameterized by $2M$ coefficients, which are the M fundamental frequencies $f_0^{(1)}, \ldots, f_0^{(M)}$ and M related inharmonicity coefficients $\beta^{(1)}, \ldots, \beta^{(M)}$. Using equations (8) and (9), a chord is thus characterized by the set $\mathcal{C} = \left\{\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(M)}, \mathcal{N}\right\}$, *i.e.* by the way how the frequency sets of the various elements of the chord are built.

Our pitch estimator relies on a weighted maximum likelihood (WML) method: for a chord $\mathcal{C}$, we calculate the weighted log-likelihood of the observed spectrum

$$\begin{aligned} L_X(\mathcal{C}) &= L_X\left(\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(M)}, \mathcal{N}\right) \\ &= \frac{1}{2M}\sum_{m=1}^{M} \ln \rho_{\mathcal{H}^{(m)}}\left(\widehat{R}_{\mathcal{H}^{(m)}}\right) + \frac{1}{2}\ln \rho_{\mathcal{N}}\left(\widehat{R}_{\mathcal{N}}\right) \end{aligned} \tag{12}$$

2.4.3 Log-likelihood normalization

For any given set of bins $\mathcal{S}$ and transfer function R, the statistical properties of the flatness $\rho_{\mathcal{S}}(R)$ depend both on the statistical properties of the whitened data X/R and on the number of observations $\#\mathcal{S}$. The latter influence tends to lower $\rho_{\mathcal{S}}(R)$ when $\#\mathcal{S}$ increases, which raises a double issue when using expression (12):

- since $\#\mathcal{N} \gg \#\mathcal{H}^{(m)}$, the various terms of the sum have values and variations that cannot be compared
- the lack of homogeneity also appears when comparing the log-likelihoods $L_X(\mathcal{C}_1)$ and $L_X(\mathcal{C}_2)$ of two chords $\mathcal{C}_1$ and $\mathcal{C}_2$ since bins are not grouped in comparable subsets.

Thus expression (12) is normalized as:

$$\begin{aligned} \tilde{L}_X(\mathcal{C}) = \frac{1}{2M}\sum_{m=1}^{M} \frac{\ln \rho_{\mathcal{H}^{(m)}}\left(\widehat{R}_{\mathcal{H}^{(m)}}\right) - \mu_{\#\mathcal{H}^{(m)}}}{\sigma_{\mathcal{H}}} \\ + \frac{1}{2}\frac{\ln \rho_{\mathcal{N}}\left(\widehat{R}_{\mathcal{N}}\right) - \mu_{\#\mathcal{N}}}{\sigma_{\mathcal{N}}} \end{aligned} \tag{13}$$

where $\mu_{\#\mathcal{S}}$ is the empirical median, depending on the observation vector size $\#\mathcal{S}$ and $\sigma_{\mathcal{S}}$ is the (constant) empirical standard deviation of $\ln \rho_{\mathcal{S}}$, both being adaptively estimated from the frame. $\tilde{L}_X(\mathcal{C})$ thus represents a normalized log-likelihood, with mean and variance around 0 and 1 respectively.

2.4.4 Silence model

When a given chord $\mathcal{C}$ is played, a peak appears such that $\tilde{L}_X(\mathcal{C}) \gg 1$, whereas for $\mathcal{C}' \neq \mathcal{C}$, $\mathcal{C}' \mapsto \tilde{L}_X(\mathcal{C}')$ is a centered process with variance 1. However, when no chord is played, no peak is generated in function $\tilde{L}_X(\mathcal{C})$ at the expected "empty chord" $\mathcal{C}_0$ = silence. A simple solution to detect silences is to set a constant value $\tilde{L}_X(\mathcal{C}_0) = \tilde{L}_0 \geq 1$ in order to obtain $\tilde{L}_X(\mathcal{C}) \gg \tilde{L}_0 = \tilde{L}_X(\mathcal{C}_0)$ when chord $\mathcal{C}$ is played, and to guarantee $\tilde{L}_X(\mathcal{C}_0) \geq \tilde{L}_X(\mathcal{C})$ for possible $\mathcal{C} \neq \mathcal{C}_0$ when no chord is played, since $\tilde{L}_X(\mathcal{C}) \leq 1$ in this case. In our implementation, $\tilde{L}_0$ is empirically set to 2.

2.5 Detection of repeated notes and MIDI file generation

The sets of notes estimated in frames of successive segments are assembled to obtain an overall discrete-time piano roll of the piece. When a note is detected both at the end of a segment and at the beginning of the following one, it may be either a repeated note, or a single one overlapping both segments. The variation of the note loudness is estimated by using the function introduced in eq. (1) for candidate selection, derived on the non-whitened version of the DFT. A note is then considered as repeated when its loudness variation is greater than a threshold empirically set to 3 dB. A MIDI file is finally generated with a constant loudness for each note.

3. EXPERIMENTAL RESULTS

Our transcription system is evaluated on a set of 30s-truncated versions of 90 piano pieces randomly chosen from B. Krueger's MIDI file database[2]. Tests are carried out using a 93-ms frame length with a note search range composed of 60 notes between C2 (MIDI note 36) and B6 (MIDI note 95). $N_c = 9$ notes are used for F_0 candidate selection, maximum polyphony is set to $P = 5$ and λ_0 is empirically set to 0.9. In order to compare the transcription with a reliable ground truth, audio files are first generated from the original MIDI files and the synthesized audio files then constitute the input of the transcription system. The audio generation is performed by virtual pianos (Steinberg The Grand 2, Native Instrument Akoustik Piano and Sampletekk Black Grand) that use a large amount of sampled sounds and provide a high quality piano synthesis. A note-based evaluation is drawn: a MIDI file comparison is performed between the original file and the transcription by classifying notes into True Positives (TP), False Positives (FP) and False Negatives (FN). TP are defined as notes with a correct pitch (with a quartertone tolerance) and an onset time error lower than 50 ms (commonly used threshold), FP are the other transcribed notes and FN are the non-transcribed notes. Performance is then evaluated through four rates: *recall* $\frac{\#\mathrm{TP}}{\#\mathrm{TP}+\#\mathrm{FN}}$, *precision* $\frac{\#\mathrm{TP}}{\#\mathrm{TP}+\#\mathrm{FP}}$, *F-measure* $\frac{2\ precision\times recall}{precision+recall}$ and *mean overlap ratio* $\frac{1}{\#\mathrm{TP}}\sum_{i\in\mathrm{TP}}\frac{\min(\mathrm{offsets}_i)-\max(\mathrm{onsets}_i)}{\max(\mathrm{offsets}_i)-\min(\mathrm{onsets}_i)}$, where onsets_i (or $\mathrm{offsets}_i$) denote the pair of onset (or offset) times for TP note i in the reference and in the transcription. The mean overlap ratio is an averaged ratio between the length of the intersection of the temporal supports of an original note and its transcription, and of the length of their union. Note that offset times are not taken into account when classifying notes between TP, FP and FN sets. This choice is due to two main reasons. First, since piano notes are damped sounds, endings are difficult to determine and depend on the instrument and on the recording conditions, even for a given original MIDI file. Secondly, the evaluation system used here enables to take offset times into account by means of the mean overlap ratio. Hence F-measure, precision and recall rates focus on the evaluation of pitches and onset times.

The results[3] of our system are reported in Figure 2, together with the performance of state-of-the-art methods [4, 5, 6]. All transcription systems are tested on the same database as our system, using their authors' original

[2] `http://www.piano-midi.de/`

[3] See also the audio examples available on the authors' web site at: `http://perso.enst.fr/~emiya/EUSIPCO08/`

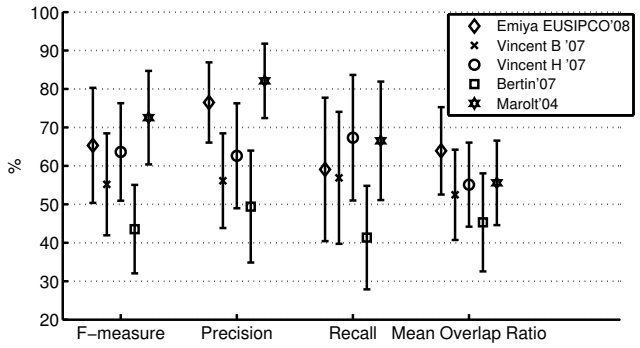

Figure 2: Evaluation results: F-measure, recall, precision and mean overlap ratio for several transcription systems, averaged on transcriptions of several pieces. *Vincent B '07* and *Vincent H '07* stand for the "Baseline" and "Harmonic" methods presented in [6]. Vertical lines show standard deviations of indicators w.r.t. all pieces, showing how results depend on music excerpts.

code. Performance rates have been averaged with respect to all test pieces, and the related standard deviation is also represented. Results dramatically depend on musical excerpts, with a number of consequences. Thus, high performance (*e.g.* F-measure greater than 90%) is reached for pieces with slow tempo or low polyphony while fast pieces are generally difficult to transcribe. This explains the large standard deviations, and means that such absolute figures should not be compared to results from other publications since they drastically depend on the database. Besides, the confidence in the results is not directly related to these standard deviations: it has been assessed by an ANOVA test on F-measure data, passed using a 0.01 test level.

Our approach is comparable to state-of-the-art systems in terms of global performance. The obtained F-measure (65%) is satisfying. We obtain a better mean overlap ratio than the competing methods, which suggests that the proposed HMM framework for note tracking is efficient both for selecting pitches among candidates, and for detecting their possible endings. It results in an efficient transcription of durations, enhancing the phrasing similarity with the original piece and thus participating in the subjective quality when hearing the correct transcribed notes. Similar results were obtained when the system was tested with real piano recordings, using 30s-excerpts of pieces used in [1]. In this case, the average F-measure reaches 79%, the musical content of the excerpts appearing quite easy to transcribe (low polyphony, medium tempi).

Generally, the main errors of our system are: polyphony limitation, missed notes in the onset detection and candidate selection stages, errors at the bass and treble bounds and harmonically related confusions (octave and others). Our system is implemented in Matlab and C, and its running time is less than 200 times realtime on a recent PC.

4. CONCLUSIONS

In this paper, we have put forward a new approach to the automatic transcription of piano music, applying some recent advances in pitch estimation and including an original structure to deal with joint estimation. The system has been tested on a database of musical pieces, compared with competing systems and has reached a satisfying performance.

Future work will deal with the estimation of loudness of notes, the question of overlap between overtones and the development of a more accurate silence model.

REFERENCES

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

[2] M. Davy, S. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *J. Acous. Soc. Amer.*, vol. 119, no. 4, pp. 2498–2517, 2006.

[3] 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, 2007.

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

[5] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Proc. of ICASSP 2007*, Honolulu, Hawaii, USA, Apr.15–20 2007, vol. I, pp. 65–68.

[6] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *Proc. Int. Conf. Audio Speech and Sig. Proces. (ICASSP)*, Las Vegas, Nevada, USA, Mar. 30 – Apr. 4 2008.

[7] G. Poliner and D. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP Journal on Advances in Signal Processing*, vol. 8, pp. 1–9, 2007.

[8] M. Ryynänen and A.P. Klapuri, "Polyphonic music transcription using note event modeling," in *Proc. IEEE Work. Appli. Sig. Proces. Audio and Acous. (WASPAA)*, New Paltz, NY, USA, Oct. 2005, pp. 319–322.

[9] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of inharmonic sounds in colored noise," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Bordeaux, France, Sept.10–15 2007, pp. 93–98.

[10] C. Raphael, "Automatic transcription of piano music," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, Paris, France, 2002.

[11] M. Alonso, G. Richard, and B. David, "Extracting note onsets from musical recordings," in *Proc. of the ICME*, Amsterdam, The Netherlands, July6–8 2005, pp. 1–4.

[12] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, Springer, 1998.

[13] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *J. Acous. Soc. Amer.*, vol. 43, no. 4, pp. 829–834, 1968.

[14] 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, 1989.

[15] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[16] R. Badeau and B. David, "Weighted maximum likelihood autoregressive and moving average spectrum modeling," in *Proc. Int. Conf. Audio Speech and Sig. Proces. (ICASSP)*, Las Vegas, Nevada, USA, Mar.30 – Apr.4 2008.