AUTOMATIC TRANSCRIPTION OF PIANO MUSIC

Valentin Emiya, Roland Badeau, Adrien Daniel, Bertrand David

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

ABSTRACT

The automatic transcription of music is a key task in the field of information retrieval in audio signals. This paper summarizes recent works on automatic transcription of piano music. The first one is a full transcription system that analyzes an input recording and provides a MIDI file including the estimation of notes. The multipitch estimation stage of the system is based on a method which is detailed separately. Finally, we report some advances in the evaluation of the resulting transcriptions in order to obtain relevant metrics from a perceptually-based point of view.

1. INTRODUCTION

Automatic transcription of music refers to the analysis of the recording of a musical piece in order to extract the part of its contents related to notes: pitches, onset times, duration, and sometimes higher-level features like rhythm patterns, key and time signatures, etc. As a process for information extraction, the automatic transcription of music is a key task in the field of Music Information Retrieval (MIR) in which it is not only a target application, making it possible to implement some audio-to-score and audio-to-MIDI algorithms, but also a basic component for other applications such as indexing and classification tasks, query by humming and other kinds of similarity analysis, or score alignment and following. While many systems have already been proposed for about thirty years [1–4], the automatic transcription of music is still a very active research field, giving rise to new approaches and more and more satisfying results.

This paper is focusing on the automatic transcription of piano music and its concomitant tasks. From a musical and a state-of-the-art point of view, while the piano is a widely-used instrument within the field of western music, while pieces for piano solo are so numerous, the piano stands among the most difficult musical instruments to be transcribed automatically (*e.g.* see [5]). This may be due to its large fundamental frequency (F_0) range and to the virtuosity of pieces for piano, causing fast and compact groups of notes and high polyphony levels, but also to intrinsic characteristics like the inharmonicity or the beats occurring in its sounds. This thus motivates our choice to investigate transcription methods specific to piano music.

This paper is structured as follows. A full system for automatic transcription of piano is introduced in Section 2. It includes a method for the estimation of simultaneous pitches detailed in Section 3. Finally, we raise the question of the evaluation of the resulting transcriptions and propose some enhancement to the usual evaluation metrics in Section 4.

The research leading to this paper was supported by the European Commission under contract FP6-027026-K-SPACE.

2. AUTOMATIC TRANSCRIPTION SYSTEM

The input of our transcription system [6] is a monaural recording of a piece of piano music sampled at 22 kHz. This signal is considered as a sum of notes and noise that is observed and analyzed in 93ms overlapping frames. Each note is modeled by a sum of sinusoids, the so-called partials, with frequencies distributed according to the inharmonicity law:

$$f_h = h f_0 \sqrt{1 + \beta h^2} \tag{1}$$

where f_0 is the fundamental frequency, which identifies the note, β is the inharmonicity coefficient and h is the partial order. This pseudo-harmonic, piano-specific distribution is due to the stiffness of piano strings: β values are typically around 10^{-3}, depending on the note whereas the spectrum of musical instruments like winds or bowed strings are perfectly harmonic (*i.e.* $\beta = 0$). Besides, two additional assumptions help characterizing piano sounds: the onsets of notes are percussive (there are no tied notes like with other instruments) and frequency modulations are not significant (no glissando, no vibrato).

In this context, we adopt the following strategy to perform the transcription task:

- Onset detection: the signal is segmented according to the apparition of new events (*i.e.* notes).
- F_0 candidate selection: after each onset, a superset of likely notes is estimated in order to lower the subsequent computations and to estimate accurate values of F_0's and inharmonicity coefficients using eq. (1).

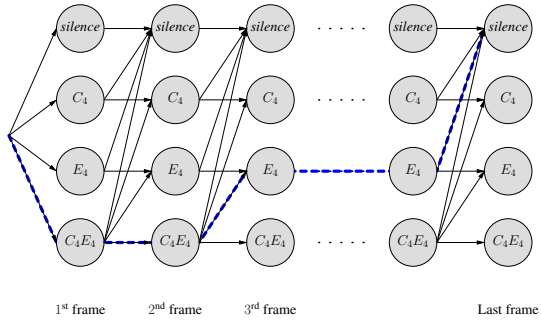

Fig. 1. Example of HMM tracking of most likely notes [6].

- HMM tracking of most likely notes: between two consecutive onsets, one hidden Markov model (HMM) is used to select the most likely path among all combinations of note candidates, each combination being a hidden state (see Fig. 1).
- a MIDI file is generated as the output of the system[1].

3. MULTIPITCH ESTIMATION

Some particular works have been dedicated to the development of the multipitch estimation method [7]. It has then been embedded in the HMM tracking of the above transcription system, as a mean to estimate the likelihood of any observed frames in any given states.

The main idea consists in finding parametric models for the spectral envelopes of notes and for the noise. By using a low-order autoregressive (AR) model, the spectral smoothness principle [8] is implemented, allowing to deal with the variability of piano spectra. Besides, the parametric aspect makes it possible to derive an estimator for the amplitudes of partials in the case of overlapping spectra. The noise is modeled by a moving-average (MA) process, which is a more fitting model for audio signals than the commonly-chosen white noise, and is more discriminating than an AR model when attempting to analyze a sinusoid+noise mixture. The resulting multipitch estimation technique is providing the likelihood of the data given a set of possible simultaneous notes and is robust for estimating the polyphony level (*i.e.* the number of simultaneous notes).

4. EVALUATION OF TRANSCRIPTIONS

Usually, transcription systems are evaluated by generating a set of transcriptions, by classifying the notes among correct estimations (or true positives, TP), false alarms (or false positives, FN) and missing notes (or false negatives, FN) and by using a metric like the so-called F-measure, defined by

$$f \triangleq \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} = \frac{\#\text{TP}}{\#\text{TP} + \frac{1}{2}\#\text{FN} + \frac{1}{2}\#\text{FP}} \quad (2)$$

in order to give a score to each algorithm. However, in that expression, each error has an equal weight. In a perception test [9], we showed that errors could be divided in several classes, each of them having its own perceptual weight. A perceptually-based version of the usual metrics can thus be defined, such as the following perceptual F-measure:

$$f_p \triangleq \frac{\#\text{TP}}{\#\text{TP} + \sum_{i=1}^{6} \alpha_i w_i \#E_i} \quad (3)$$

with perceptual weights α_i for typical errors i (octave, fifth, other intervals, deletion, duration, onset) and time-related errors w_i.

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

References

[1] J.A. Moorer, *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer*, Dept. of Music, Stanford University, 1975.

[2] A.T. Cemgil, *Bayesian Music Transcription*, Ph.D. thesis, SNN, Radboud University Nijmegen, the Netherlands, 2004.

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

[4] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*, Springer, 2006.

[5] G. Peeters, "Music pitch representation by periodicity measures based on combined temporal and spectral representations," in *Proc. of ICASSP 2006*, Toulouse, France, May 2006, vol. 5, pp. 53–56.

[6] V. Emiya, R. Badeau, and B. David, "Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches," in *Proc. of EUSIPCO*, Lausanne, Switzerland, Aug. 2008.

[7] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of inharmonic sounds in colored noise," in *Proc. of DAFx*, Bordeaux, France, Sept. 2007, pp. 93–98.

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

[9] A. Daniel, V. Emiya, and B. David, "Perceptually-based evaluation of the errors usually made when automatically transcribing music," in *Proc. of ISMIR*, Philadelphia, Pennsylvania, USA, Sept. 2008.