Automatic Transcription of Monophonic Audio to MIDI

Jiří Vass¹ and Hadas Ofir²

¹ Czech Technical University in Prague, Faculty of Electrical Engineering
Department of Measurement
vassj@fel.cvut.cz

² Technion - Israel Institute of Technology, Department of Electrical Engineering
Signal and Image Processing Laboratory
hadaso@siglab.technion.ac.il

Abstract This paper presents an automatic system for transcription of monophonic audio signals into the MIDI (Musical Instrument Digital Interface) representation. The system incorporates two separate algorithms in order to extract the necessary musical information from the audio signal. The detection of the fundamental frequency is based on a pattern recognition method applied on the constant Q spectral transform. The onset detection is achieved by a sequential algorithm based on computing a statistical distance measure between two autoregressive models. The results of both algorithms are combined by heuristic rules eliminating the transcription errors in a natural manner. The method is suitable for rapid musical passages, able to deal with various musical sounds, and applicable within a wide range of MIDI frequencies.

1 Introduction

Transcription of music is a task of converting a particular piece of music into a symbolic representation, such as standard musical notation or MIDI file format. From this point of view, music can be classified as polyphonic and monophonic. The former consists of multiple simultaneously sounding notes, whereas the latter contains only a single note at each time instant, such as a saxophone solo. Monophonic transcription can thus be understood as a special simple case of polyphonic transcription, and is considered as practically solved [1]. On the other hand, it represents an important case to be treated separately with much stricter demands on the transcription quality (which still seems to be relatively limited for polyphonic transcribers). Moreover, specific applications for monophonic systems comprise tools for solo musicians [3], low bit-rate audio coding [4], and monophonic ring tones for cellular phones.
Each musical note can be described by three essential parameters: the fundamental frequency (pitch), the beginning of a note (onset time), and the note duration. For this reason, a transcription system should include both pitch tracker and onset detector, although not necessarily implemented as two separate blocks. Previous papers tend to describe techniques for pitch and onset detection rather separately, so only few compact and reliable monophonic transcribers were published (for example [3]). Furthermore, transcribers based on autocorrelation [4] suffer the common time-frequency tradeoff, and cannot be applied to a wide range of frequencies [2], which is an essential property of audio signals. Our solution is therefore based on the Constant Q Transform (CQT) offering adjustable frequency range and excellent time-frequency resolution, which is further improved using an onset detector on a sample-by-sample basis.

2 The Transcription System

Figure 1 depicts the basic building blocks of the transcription system.

![Figure 1. Transcription System](image)

The Pre-processing block normalizes the input signal to the unit power, and inserts silence before the beginning and after the end of the signal. The Pitch Detection block is primarily responsible for tracking the fundamental frequency in the signal, but also contributes to the determination of the note onsets and offsets, which is the principal task of the Detection of Events block (whose output can conversely affect the
fundamental frequency tracking). Since none of these blocks yield ideal results, the Estimation of Power block is added to provide supportive data for event detection. The Combining the Results block then processes the outputs of the preceding sections, and generates the complete event list, which is finally converted to a MIDI file by the Notes to MIDI block developed by [2].

2.1 Pitch Detection

Since the musical frequencies form a geometric series, it is desirable to represent the signal with a spectral transform corresponding to a filter bank with center frequencies spaced exponentially. Such transform was developed by [5] and is referred to as Constant Q Transform (CQT). Similarly as in the DFT, the frequency range is divided into frequency bins, each represented by a bandpass filter with a center frequency $f_k$ and filter bandwidth $\Delta f_k$. However, the CQT bins are geometrically spaced, which results in variable resolution at different octaves and constant quality factor $Q = f_k/\Delta f_k$. With $b$ being the number of filters per octave, the CQT filter bank is equivalent to $1/b^{th}$ octave filter bank, which shows its relationship with the wavelet transform [2]. The CQT spectral components form a constant pattern when plotted against the logarithmic frequency axis, which is evident from the following equation:

$$\log (f_m) - \log (f_n) = \log \left( \frac{f_m}{f_n} \right) = \log \left( \frac{m \cdot F_0}{n \cdot F_0} \right) = \log \frac{m}{n} \quad (1)$$

In other words, the relative positions between harmonics are constant for all musical notes, and only the absolute positions depend on the fundamental frequency $F_0$. This property can be employed to determine the fundamental frequency as a maximum value of the cross-correlation function between an ideal (theoretical) pattern and the pattern in the actual CQT spectrum [6], as depicted on Fig.2.

2.2 Detection Of Events

The algorithm for detection of acoustic changes (events) is based on [7], and was also successfully applied on audio signals [8]. Great advantage is the statistical time-domain approach performed on a sample-by-sample
basis, thus providing very accurate locations of the onset and offset times. The main idea is to model the signal by two autoregressive (AR) models, between which a suitable distance measure is monitored, and a new event detected whenever the distance exceeds a specific threshold value. Since the distance measure is *conditional Kullback's divergence*, the algorithm is commonly referred to as the *divergence test*.

### 2.3 Estimation of Power

Since the divergence test provides no information about the origin of the detected acoustic event, it is necessary to estimate the signal power in order to reliably distinguish between the true and false onsets. This is achieved using the *leaky integrator* which computes the recursive estimate of power [9]. The resulting power signal is then iteratively smoothed to locate its most significant minima, which are subsequently utilized in the final decision procedure.

### 2.4 Combining the Results

This block processes the outputs of the preceding sections and applies several heuristic rules to correctly choose the best candidate for the onset time and MIDI frequency of each note. These rules form an eliminative competition between two sets of candidates obtained by the CQT and the divergence test (shown on Fig.3).

**Rule 1** The first rule can briefly be summarized as: "The winner is the nearest". Specifically, the algorithm sequentially processes the candidates from the CQT segmentation, and assigns to each member the nearest candidate from the divergence test.
Rule 2  The correct candidates from the divergence test are typically located in the vicinity of power minima preceding the note attacks (see Sec.2.3). This property is thus the necessary condition of the second rule, which allows an additional acceptance of the divergence test candidates rejected by Rule 1. This rule enables detection of successive notes of the same frequency, and hence allows the time segmentation which would not be possible solely with the CQT approach.

Rule 3  The third rule is based on the observation that monophonic signals contain an attack between each two consecutive onsets. In other words, the power must have at least one local maximum to consider such onsets as beginnings of two separate notes.

![Figure 3. Results of CQT and the divergence test to be processed by heuristic rules](image-url)
3 Conclusion

This paper presents a compact system for transcription of monophonic audio signals to MIDI representation. The time-frequency resolution of the pitch tracker is improved by a sequential onset detector, thus resulting in accurate transcription both in time and frequency. Moreover, the new method applies several heuristic rules to eliminate the errors of both pitch- and onset-detection algorithms. Such approach avoids the conventional parameter of minimum note duration, and is therefore especially suitable for transcription of fast musical passages.

References