Pitch-to-MIDI conversion

Jari Salo

10.11.2003

Abstract

There are several kinds of methods and underlying models for monophonic fundamental frequency estimation of periodic or semi-periodic signals. The methods operate in various domains (time-, frequency- or joint time-frequency domains). A general structure and few examples of different approaches and their strengths as well as weaknesses are briefly described.

1 Introduction

Digital audio data, including multimedia applications, has grown rapidly during the recent years. This means that databases that have traditionally dealt only with text and numeric data are now increasingly faced with digital representations of music. Text-or SQL based queries have become insufficient for indexing and querying. They are being replaced by content-based retrieval (CBR) that is using methods natural to the new data types in order to enable intelligent navigation and search. For example an easy way for searching a library of songs would be whistling, singing or humming a short fragment of a song into a microphone. As melody is one of the most prominent features of music, this recorded fragment would allow the system to retrieve closest matches from the database.

The subject of this study is an automatic transcription of a recorded (monophonic) melody into a musical instruments digital information (MIDI) representation. MIDI information, in its most basic mode, tells when to start and stop playing a specific note. Other information shared includes the volume and modulation of the note, if any. From the point of view of information retrieval (IR) the relevance of such transcription is very straightforward as MIDI provides an extremely compact cartoon-like representation of musical data. Yet as much of the necessary information of the contents of the audio recording is retained, the resulting metadata can be used for many contents based analyses and queries.

2 The general approach

The majority of automatic transcription systems concentrate on analyzing and processing a recorded human voice. Human voice is monophonic, and this means that there is only one fundamental frequency (F0, perceived as pitch) at a time. However as voice can have great variations in range, intensity, expression and timbre (the quality given to a sound by its overtones), it is the most natural and a most challenging musical instrument to study.

A common structure for the transcription systems is to have two or three different processing stages:

- The first stage, pre-processing, is dealing with the problems caused by the actual recording: acoustic interference, background sounds, hum, buzz and noise in the recording chain and also the problems caused by the analog to digital-conversion.

- The second stage, processing, identifies the pitch-contour: each musical event with a pitch, onset and offset position. Also the loudness (intensity) of the event is sometimes detected. This stage has it's own challenges, among them are vibrato (slight and rapid variations in pitch), glissando (a sliding up or down the musical scale), tremolo (rapid variations in loudness) and separating consonants that are high-frequency content noise-like segments from actual noise.

- The final stage, post-processing, enhances the received results and labels notes to the detected pitches. This stage will deal with interpreting the information derived from the previous two stages to correct errors. This includes adjusting the tuning of the musical scale of the singer to the standard tuning that MIDI uses.

3 Noise reduction and other preprocessing

Even if an audio recording is made in a soundproof booth, it will contain at least some additive noise from the environmental. In addition it will probably also contain convoluted noise due to changing room acoustics and resonance, changes in microphone or other parts of the recording chain etc. While humans have the ability to extract and listen to harmonic signals even in very noisy environment [1], the accuracy of pitch prediction algorithms decreases rapidly when noise level goes up. To make things worse, the spectra of real environmental noises are in general not white but colored [2]. Thus the main tasks of the preprocessing stage are noise reduction and enhancing the features that are useful for fundamental frequency estimation.

4 Fundamental frequency detection

Pitch detection is of interest whenever a single semi-periodic sound source is to be examined or modeled, specifically in speech and music. In fact many techniques to extract pitch information from audio signals have been primarily developed for speech and then extended to the music domain. [3] Regardless of the application, a pitch tracker should ideally have a stable and accurate output, a good tolerance for noise, have no significant delay and work on any monophonic instrument. In contrast to speech signals, musical instruments can have an extremely large frequency range, and a multipurpose pitch-tracking algorithm has to deal with a very large bandwidth. [4] Though in theory the detection of pitch from monophonic sources is well understood, finding the pitch of musically interesting sounds is extremely difficult, especially under the above mentioned constraints. In order to minimize the delay the pitch must be determined at the very beginning of the note. This part is called "attack" and it is an important aspect of defining a particular timbre of an instrument. [5] It can be harmonically very complex and also contain inharmonic partials. Also the sustaining part of the notes can vary constantly without onset of notes. Inexplicable pitches and unpitched sounds provide even more challenges [4].

The methods for detecting the pitch can be divided roughly into two categories: operating either in time domain or frequency domain. However the distinction between time-domain and frequency-domain algorithms is not always so clear as some of the algorithms can be expressed in

both time-and frequency domains and some systems might use multiple algorithms, possibly working in both domains.

One of the first and simplest techniques toward estimating the frequency content of a signal is zero-crossing rate (ZCR), consisting of counting the number of times the signal crosses the 0-level reference in order to estimate the signal period. The value of ZCR has also been found to correlate strongly with spectral centroid, also called spectral balancing point, which is the first moment of spectral power distribution. [6] It follows that the value of ZCR has more to do with timbre than with pitch.

Autocorrelation is maybe the most common time domain method, and the average magnitude difference function is basically same as autocorrelation but faster to implement in integer arithmetic. They are based on detecting similarity between the waveform and a time lagged version of itself. These functions may contain too much information, most of which is not related to the fundamental frequency, and that's why the signal is usually pre-processed to make the periodicity more prominent and to suppress other distracting features. Such techniques are often called "spectrum flattening".

The other main group of methods operates in the frequency domain. This type of analysis is concerned with the decomposition of the observed series into periodic components. The main analytical tool for spectral analysis is the spectrum of a series of harmonic frequencies, called *Fourier frequencies*. Almost all systems start by dividing the input signal into partly overlapping and windowed frames and then the short-time spectrum of the frame is obtained by taking a discrete Fourier transform (DFT). Windowing of the signal is recommended to avoid spectral smearing, and depending on the type of window, a minimum number of periods of the signal must be analyzed to enable accurate location of harmonic peaks. [7] With the prominent spectral peaks detected , their parameters, amplitudes, frequencies, and phases, can be estimated [8]. The common divisor of the harmonic series that best explains the spectral peaks is the fundamental frequency [3]. Various linear preprocessing steps can be used to make the process of locating frequency domain features easier, such as performing linear prediction on the signal and using the residual signal for pitch detection [9]. Performing non-linear operations such as peak limiting also simplifies the location of harmonics.

A natural way to expand the basic frequency domain analysis is to use a multi-band approach. Independent pitch estimates are calculated at separate frequency bands, and then the results are combined to yield a global estimate. This solves several problems, one of which is inharmonicity. In inharmonic sounds, as stretched strings, the higher harmonics may deviate from their expected spectral positions, and even the intervals between them are not constant. However the spectral intervals can be assumed to be piece-wise constant at narrow enough. Another advantage of bandwise processing is that it provides robustness in the case of badly corrupted signals, where only a fragment of the whole frequency range is good enough to be used. [10]

Wavelet Transform (WT) is a non-parametric analysis tool, which allows localization both in time and in frequency domains. The result of a WT is analogous to Fourier Transform performed at every time point of the time series and resulting in a frequency spectrum generated for every time moment. The main difference between Short Term Fourier Transform (STFT) and WT is that STFT is a constant-bandwidth analysis method, whereas WT is a constant-Q analysis method, which resembles auditory filters. Wavelet coefficients are obtained by computing the correlation between each wavelet and the signal. The realizable form of continuous wavelet transform is called sampled CWT (SCWT), which is most widely used in speech signals analysis [11], but can also be adapted to music analysis. The basic pitch detection idea is to filter the

signal using a wavelet with derivative properties. The output of this filter will have maxima where zero crossings happen in the input signal. After detection of these maxima, the fundamental frequency will be estimated as the distance between consecutive maxima. This filtering function will combine the bandwidth properties of the wavelet transform at different scales. [12]

Although frequency-domain algorithms may yield higher accuracy, time-domain algorithms have the advantage that they can be implemented with minimal difficulty on a general-purpose computer. This is especially important when designing systems that need to work in real time, with minimum output latency and/or for hardware with low processing power.

5 Pitched/unpitched decision

The portions of a sound that cannot be well represented with a harmonic model are generally considered as noise. Zero crossing, average energy and harmonic distortion are typical tools for discriminating between pitched and unpitched segments. However the similarities of unpitched vocal and instrument sounds to noise make task of separating the two a very complicate one. Pitch values can be useful to improve signal segmentation through the detection of noise-like segments, silences and percussive elements. For example because the pitch range of human voices is approximately between 60 Hz and 1200 Hz, each pitch detection frame that exceeds these limits could either be silence or noise, a consonant (such as /s/) or an error. If a number of adjacent frames exhibit the same noise-like behavior, they identify a wide region of non-vocal signal. On the other hand a typical spectrum of an unvoiced (no vocal cord vibration) event is characterized by (usually strong) high frequency components with a duration that is consistently shorter than voiced segments (5-10 msec for unvoiced; 50-100 msec for voiced). [3]

6 Segmentation

In a tune the basic informative unit is note. The segmentation process aims to divide the audio signal into notes. The major problem here is the difficulty in defining a representation suitable for the task. Segmentation solely based on energy will usually fail in detecting notes in a signal that exhibits a high level of variability. Unlike musical instrument timbre, human voice does not show a typical pattern that represents, for example, the onset of a tone. Moreover, sound level and source spectrum change in relation to pitch, the distance of pitch from the formants (harmonics) and the sequence of phonemes to be sung. [3]

By inspecting the shape of the envelope of sounds, it is revealed that deep valleys are valid indicators of note changes with two exceptions: events separated by a pause and notes played or sung “legato”. Therefore, a feature for signal/silence detection combined with pitch and envelope information can be used as a basic segmenter. Second, a finer segmentation is achieved by studying the shape of the envelopes in order to estimate local maxima and minima. [3]

7 Post processing

The fundamental frequency contour that is the output of the different pitch detection algorithms can be badly affected with isolated errors, and this is why different methods for correcting them have been defined. A very simple approach (similar to pitched/unpitched decision) is to define certain thresholds for either absolute or relative value of the detected pitches and then remove the pitches that cross the threshold. [13]

Another method is to smooth a function is the convolution of the input signal with the impulse response of a low-pass filter. Since the smoothing function (window) usually is of very short length, this convolution can be reduced to the weighted addition of few samples. Since the convolution is linear, we speak of linear smoothing. The application of low pass filters removes much of the local jitter and noise, but it does not remove local gross measurement errors, and, in addition, it smears the intended discontinuities at the voiced-unvoiced transitions. [12] However, filtering the output period can also be done with some kind of nonlinear smoothing filter, or median filter that can further improve measurements since spurious peaks and overshoots in the pitch estimation process are filtered out. [14] A combination of median and linear smoothing filters can also work well, as the median removes short errors, and the linear smoothing removes jitter and noise [15].

8 Scaling

Musically untrained people exhibit usually only a vestige of absolute pitch but are demonstrated to be reasonably accurate in singing intervals. Therefore, the adoption of an equal tempered musical scale for note labeling without modifications could be a source of further errors. Nonetheless, the precision in humming, singing etc. intervals indicates the equal temperament as the right context in which evaluating pitches. Therefore, a musical scale relative to the performer has to be embedded in systems that aim to translate the produced melody into music. The main advantage of a similar approach is the minimization of the error in the labeling stage. The precision of a query-by-humming or similar system depends on the understanding of the query, so that a nearly perfect translation of the input at the semitone level remains an important goal. [3]

9 Conclusion

Even after decades of work and a great number of different theories and working models, robust pitch estimation remains an intriguing task. There is no proof of any system that will always outperform others in noisy and/or otherwise difficult conditions. Also finer details like accurate voiced/unvoiced detection, or tracking of varying pitch are not yet well defined. So far most of the systems compare the achieved results to those manually transcribed by professional musicians, thus "perfect" pitch detection itself is also an undetermined problem.

References

[1] Barker, Jon & Cooke, Martin & Ellis, Dan. (2000*). "Decoding Speech In The Presence Of Other Sound Sources"*. ICSLP00, Beijing, China.

[2] Byun, Kyung Jin & Jeong, Sangbae & Kim, Hoi Rin, & Hahn, Minsoo. (2003). "Noise Whitening-Based Pitch Detection for Speech Highly Corrupted by Colored Noise". *ETRI Journal* Vol. 25, No. 1, Feb. 2003.

[3] Pollastri, Emanuele. (2002-2003). *"Processing Singing Voice for Music Retrieval"*. Ph.D. thesis, Universita Degli Studi Di Milano.

[4] Jehan, Tristan. (1997). "*Musical Signal Parameter Estimation*". Master of Science Thesis, Center for New Music and Audio Technologies, University of California, Berkeley.

[5] Garcia, Mauricio Freire. (2001). "Density 21.5 by Edgard Varèse". *Mikropolyphonie - The Online Contemporary Music Journal,* Volume 7, 2001.

[6] Panagiotakis, C. & Tziritas, G. (2003). "*A Speech/Music Discriminator Based on RMS and Zero-Crossings*". IEEE Transactions on Multimedia, 2003

[7] Charif, R. A., Mitchell, S. & Clark, C. W. (1995). "*Canary 1.2 User's Manual*". Cornell Laboratory of Ornithology, Ithaca, NY, USA.

[8] Virtanen, Tuomas. (2000). "*Audio Signal Modeling With Sinusoids Plus Noise*". Master of Science Thesis, Tampere University of Technology.

[9] Li, Chunyan & Gersho, Allen & Cuperman, Vladimir. (1999). "*Analysis-By-Synthesis Low-Rate Multimode Harmonic Speech Coding"*. Eurospeech'99, Budapest, Hungary.

[10] Klapuri A. (2000). "*Qualitative And Quantitative Aspects In The Design Of Periodicity Estimation Algorithms*", EUSIPCO Conference Proceedings, 2000.

[11] T. Tan, Beng & Fu, Minyue & Spray, Andrew & Dermody, Phillip. (1996). "*The Use Of Wavelet Transforms In Phoneme Recognition*". The Fourth International Conference on Spoken Language Processing (ICSLP) , Philadelphia, 1996.

[12] Gómez, E. (2002). "*Melodic Description of Audio Signals for Music Content Processing*". Doctoral Pre-Thesis Work. UPF. Barcelona.

[13] Lee, I-Yang & Jang, J.-S. Roger & Hsu, Wen-Hao. (1999). "*Content-based Music Retrieval from Acoustic Input*". 12th IPPR Conference on Computer Vision, Graphics, and Image Processing, PP. 325-330, Taiwan, August 1999.

[14] Weihs, C. & Ligges, U. (2003). "Automatic Transcription of Singing Performances". *Bulletin of the International Statistical Institute, 54th Session*, Proceedings, Volume LX, Book 2, 507-510.

[15] Bagshaw, P. C. & Hiller, S. M. & Jack, M. A. (1993). "Enhanced Pitch Tracking And The Processing Of F0 Contours For Computer And Intonation Teaching,*". Proc. European Conf. on Speech Comm.* (Eurospeech), pp. 1003-1006.