Tracking melody in polyphonic audio. MIREX 2008

Pablo Cancela
1015, Magallanes St.
Montevideo, Uruguay
pcancela@gmail.com

Abstract

In this work a melody extraction technique is introduced to the MIREX 2008 campaign. The task's objective consists in estimating the pitch of the main melody in polyphonic audio. The proposed method is based in three separate modules. The first one evaluates the presence of quasi-stationary harmonic sources in a frame level based approach. Next, sources are tracked using the output of the first step, obtaining a set of pitch contours. In the last step, it is decided whether or not each contour at each time is part of the main melody.

Keywords: ISMIR, Multiresolution spectrum, tracking pitch, main melody.

1. Introduction

The melody of a musical piece is usually composed of stationary or quasi stationary harmonic sinusoids [1]. The presence of other audio sources in polyphonic audio introduces two main difficulties in its study. The first one is to avoid the interference of one source in the analysis of another source. This is very difficult to achieve studying each sinusoid on its own, so it is necessary to consider the harmonic coherence of all the partials of every source. While some of the partials might be nonexistent and others might be destroyed by other sources, some of them will still be locally salient in frequency and salient as a group due to their harmonic structure.

The second main problem consists in determining which of the analyzed sources corresponds to the main melody of a piece. It can be considered that the main melody is the pitch that can be naturally followed by a human. On one hand, this problem might not be well defined as this criteria is very subjective. On the other hand, imposing a model of what are typical melody properties would give reasonable results of what is understood of a main melody. But this model will rely in cultural aspects and is far from general.

The MIREX 2004 and 2005 databases are an excellent training set for this task as they cover a wide range of styles. On some of the songs there is more than one melody and ground truth follows in all of them the voice, if there is a singing voice, or the melody of main instrument if there is no singing voice.

2. Feature extraction

The first part of the algorithm is a frame level based spectral analysis to estimate the harmonic audio sources. Accurate frequency candidates at this stage are strongly necessary to be able to follow sources along time. The method tries to find locally salient frequency components, and later groups of them following an harmonic structure.

2.1. Spectral Analysis

The analysis frame length is 92,9ms (2048 samples at 22050 Hz), and the hop time is 3.4ms (75 samples at 22050 Hz). As the frame length is too long to analyze non-stationary components, a modified STFT is computed to have almost constant resolution in all frequency range for different slopes of varying frequency components. This provides a different window width for each frequency bin, and the ability to analyze non stationary components for different slow frequency changes. See [2] on this topic.

This frequency analysis permits to observe with reasonable detail slowly changing partials of any frequency at each frame. The chosen hop time is quite small, giving a very high resolution spectrogram in time, which makes the process rather slow but will be necessary to correctly track relatively fast changing fundamental frequencies.

2.2. Harmonic structures

Now, salient frequency components are enhanced. This is done high pass filtering the spectrum, killing the base line of frequency components. Before this filtering the logarithm of the absolute value (plus an offset) of the spectrum is applied to normalize power, giving a kind of cepstrum. The purpose of this is to attenuate the values of very high power components, as we will be interested in several salient harmonic peaks (and not only one with high power).

Next, an harmonicity map, let's call it $f_0 - gram$, for each f_0 in the range from 40Hz to 1280Hz is constructed. The candidate harmonic components are calculated by summing values in each calculated spectrum at integer multiples of the considered f_0: $k.f_0$. See [4].

This gives a good candidate map for each time and frequency of harmonic sources. But it also gives high response in fundamental frequencies that are small integer fractions

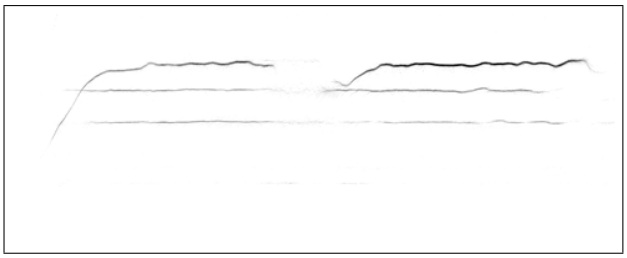

Figure 1. Example of $f_0 - gram$ for the low frequencies of the first seconds of the "pop1.wav" excerpt. Three simultaneous voices singing "Michelle" can be distinguished. The main melody is the one with higher pitch.

of an existent fundamental frequency. For example if there is an harmonic source of frequency f_0, there will also be a high sum value for fundamental frequencies: $f_0/2$, $2f_0/3$, $f_0/3$. So the sum of frequencies for a given f_0 is attenuated if the mean of components at $2kf_0$, $3kf_0/2$ and $3kf_0$ is above the mean considering all partials at $k.f_0$.

There will be also high responses at integer multiples of f_0, but the sum computed for them will miss at least half of the partials (odd partials for $2f_0$, not multiple of 3 partials for $3f_0$). It is not necessary to attenuate them, as they take always much lower values than for the actual fundamental frequency.

After this processing a $f_0 - gram$ is obtained, with very high resolution in time and frequency. This map will be the sole input of the module that tracks and analyzes sources.

3. Tracking sources

To determine the melody two steps are followed, the first one is to extract harmonic notes tracking them in the $f_0 - gram$. This is done until a minimum 'power' threshold is reached.

Time and frequency for which the highest value in the $f_0 - gram$ is chosen. The fundamental frequency is tracked from this point to the future and past in the $f_0 - gram$. The high resolution in time permits an accurate tracking for strong enough sources. This tracking is done until the magnitude of the $f_0 - gram$ in the ends of this tracked pitch contour drops below a threshold.

After the tracking, this source is erased form the $f_0 - gram$, so that it will not be considered again.

This process is repeated until the remaining maximum power is below a threshold. No more than 4 sources are allowed at the same time, but this is just for efficiency and considering more does not affect results.

At the end of this process a set of tracked pitch contours is obtained. For these contours it is known the magnitude in the $f_0 - gram$ at each time and frequency. This information will be used later to decide if this source is part of the main melody.

This set will be the input for determining the melody, the $f_0 - gram$ is not longer used in the final step.

4. Determining the Melody

In this last module a higher level decision of which pitch contours belong to the main melody is taken.

The considered criteria are: Weight less pitches with very low fundamental frequencies (below 150Hz). Weight more fundamental frequencies in which the derivative of pitch is high, to enforce singing voice that typically has a non stable changing f_0. Low pass filtering of the magnitude of one source along the pitch contour.

Then, for each instant the pitch corresponding to the maximum value is chosen. To decide voiced/unvoiced instants, an adaptive threshold is calculated to eliminate weak harmonic sources.

Finally, a mean melody pitch line is estimated, as a low pass filtering of the calculated main melody. This is used as a refinement to penalize sources that are far away from this mean pitch.

The main melody is recomputed taking all the previous considerations and adding a penalization according to the distance to the calculated mean melody pitch line.

5. Implementation and results

The method was implemented in MatLab, with C and Assembler MEX functions for repeated time consuming operations, as an extremely exhaustive search of harmonic sources is performed. The algorithm is rather slow, $60x$ real time in a Pentium 4 @1.8GHz, working with audio at a sampling rate of $22050Hz$. Doing the analysis less exhaustive gives lower processing times (10x) and reasonable good results, but times are still far from real time.

When the main melody is relatively strong and clear, results are good. There are some problems when extremely strong percussion attacks appear. Performance also goes down when there is a secondary relatively strong melody, as when the the main melody is not present the algorithm follows the second one.

6. What's missing. Future work.

Timbre is not considered in the spectral analysis, it is reasonable to assume that one harmonic source, such as a singing voice or an instrument have partials whose envelopes change slowly in time, and so does timbre. So weighting the sum of partials proportionaly to their expected power would give a better response for harmonic sources, diminishing the interference of other sources in the analysis.

High power attacks are also a problem for the method as is, highly diminishing the detection of harmonic sources. Interpolation in these places would probably improve results.

Speed! Even some parts of the algorithm are highly optimized it should be possible to improve processing time porting it completely to C. Several parts of the processing consist in lots and lots of highly parallelizable simple operations which seem ideal to be done in a GPU.

7. Acknowledgements

Thanks to the IMIRSEL crew at the University of Illinois at Urbana-Champaign for running MIREX. Thanks to Ernesto López and Martín Rocamora for their support and useful discussions.

References

[1] X. Serra. "A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition. Ph.D." Dissertation, Stanford University, 1989.

[2] J. C. Brown. "Calculation of a constant Q spectral transform." J. Acoust. Soc. Am., 89(1): 425-434, 1991.

[3] DikJ. Hermes, "Measurement of pitch by subharmonic summation", The Journal of the Acoustical Society of America, Volume 83, Issue 1, pp.257-264, 1988.