# EXTRACTION OF THE MELODY PITCH CONTOUR FROM POLYPHONIC AUDIO

**Karin Dressler**
Fraunhofer Institute for Digital Media Technology
Langewiesener Str. 22
98693 Ilmenau, Germany
dresslkn@idmt.fraunhofer.de

## ABSTRACT

MIREX 2005 is the second evaluation of algorithms related to music information retrieval (MIR). This document describes our submission to the MIREX audio melody extraction contest addressing the task of identifying the melody pitch contour from polyphonic musical audio. We use mainly a data-driven approach – implementing standard audio signal processing techniques like Fourier analysis, instantaneous frequency estimation and sinusoidal extraction. Nevertheless, some high level knowledge is applied. Examples of such knowledge include perceptually motivated methods (psychoacoustics, auditory stream segregation) as well as basic voice-leading principles. The result of the Mirex 2005 evaluation proves that our algorithm performs best in respect of runtime and overall accuracy, a measure which combines voicing detection (discrimination between melody and non-melody parts) and pitch estimation.

**Keywords:** MIREX, audio melody extraction, predominant pitch contour

## 1 INTRODUCTION

The transcription of the melody out of polyphonic audio data is recognised as an unsolved problem among researchers. While humans easily spot the melody line from various musical pieces, there is no reliable technical method for the automatic extraction of the melody.

Yet, over the last 10 years there has been a remarkable progress in the area of transcription of polyphonic music - a topic that is closely related to melody recognition. The latest publications show promising results in the transcription of polyphonic audio pieces, though in most cases the audio test data is subject to several restrictions: be it that only one special instrument is allowed, be it the limitation of the maximum number of instruments playing simultaneously or the exclusion of percussive instruments.

Particular attention received the publications of Masataka Goto introducing his now famous PreFEst (predominant F0 estimation) algorithm (Goto and Hayamizu, 1999; Goto, 2000). Goto was the first person to achieve an extraction of the melody and bass line from real world CD recordings. Focussing on the estimation of predominant musical voices rather than aiming at the transcription of all sound sources made his transcription system independent of prior information. Thus, with his approach he proved the possibility to overcome the above mentioned limitations.

The aim of the presented algorithm is the automatic extraction of the melody (predominant voice) from contemporary western music. Most scientists comply with the statement, that this difficult task cannot be solved using standard signal processing methods alone. In fact, knowledge has to be employed in the music theory or even knowledge about the processes in the auditory cortex of humans. Hereby, combining and developing further techniques from audio signal processing on and computational musicology is the main focus of our research. While emphasis is placed on algorithm efficiency rather than on modelling the human auditory system, functionality still remains the main quality criterium.

## 2 METHOD

### 2.1 Spectral Analysis

A multi resolution spectrogram representation is obtained from the audio signal by calculating the Short-Term Fourier Transform (STFT) with different factors of zero padding using a Hann window. Thereby we make use of the very efficient implementation of a Real FFT provided by Ooura (2001). For all spectral resolutions – assuming audio data sampled at 44.1 kHz – the resulting STFT frame size and the hop size of the analysis window are 2048 and 256 samples, respectively. This processing step is followed by the computation of the magnitude and phase spectra.

To gain a better frequency discrimination, the instantaneous frequency (IF) is estimated from successive phase spectra. We apply the well-known phase vocoder method proposed by Flanagan and Golden (1966) for the IF extraction.

### 2.2 Sinusoidal Modelling

Since sinusoidal components of the audio signal contain the most relevant information about the melody, a sinusoids plus noise analysis is performed on the spectral data (Serra, 1989). The underlying idea of this technique is, that an audio signal can be divided into stable partials originating from periodic sound and a noise component. Only

partials which (probably) result from a deterministic signal are used for further analysis; stochastic components are neglected.

Although temporal continuity of the sinusoidal trajectories is an important criterium for identifying eligible spectral peaks as sinusoidals in Serra's approach, this continuity is not required by our method. Charpentier (1986) found that you can identify sinusoidals by distinct spectral features in one frame alone. We developed his method further and this way improved the performance of the adjacent pitch estimation noticeably. Despite the fact that our sinusoidal detection is quiet robust against noise and distortion, it is not adequate for audio signals with a very dense spectrum, which for example can be found in many pieces of rock music.

### 2.3 Pitch Estimation

The magnitude and instantaneous frequency of the sinusoids are evaluated by a pitch estimation method, as the frequency of the strongest partial may not be the perceived pitch of a periodic complex tone. At first, the pitch estimator performs a perceptually motivated magnitude weighting and then it analyzes the harmonic structure of the signal. The algorithm covers four octaves – looking for dominant virtual pitches in the frequency range between 80 Hz and 1280 Hz.

### 2.4 Streaming

The frame-wise estimated pitch events shall now be grouped into acoustic streams. There is evidence that the auditory system uses primitive grouping rules, like pitch magnitude, pitch proximity and timbre, as well as schema-based grouping rules, for example tuning, scales and rhythm. Since schema-based mental organisation rules depend strongly on the cultural background of the music listener and moreover vary drastically among different music styles, they can hardly be applied without a significant loss of generality. Due to the fact that humans are able to identify the predominant voice even from unfamiliar, foreign music, we believe that in most cases the identification of the melody line is possible without conscious effort and prior learning. Thus we solely use cues like the continuity of pitch contours, the proximity of different contours and the computed pitch magnitude for stream building from successive pitch events. With this strategy we follow the reasoning of Albert S. Bregman, who describes different aspects of auditory stream segregation in (Bregman, 1994).

### 2.5 Identification of the Melody Stream

The most salient streams are examined further helping to identify if a stream object belongs to the melody voice or not. This challenging task can be named one of the key problems for the automatic melody retrieval from polyphonic audio. Even for pieces with a very prominent melody voice the distinction between melody and accompaniment (or noise) cannot be achieved easily. Since there is – unfortunately – no reliable method for the extraction of timbre from a mixture of simultaneously playing instruments or voices, we cannot use this certainly valuable feature as a criterion. Hence the duration of the continuous stream, which may consist of several notes, its estimated magnitude and the frequency region remain as the main decision criteria.

Those features are now observed over a longer period. We use a time window of maximum 4 seconds, which roughly complies with the capacity of the human short term memory (Snyder, 2000). All stream objects lasting longer than 100 ms are sorted according to their similarity of frequency range and stored in different registers. Doing so, we are able to identify the most active frequency regions. All stream objects (even with a very short duration), which belong to this region, gain an additional weight in the concluding comparison.

Finally a rule-based decision method chooses the melody "notes"[1] from the remaining candidates: Tone successions containing intervals larger than the octave are avoided. If there are more than two detected voices, the outer voices achieve a higher rating, because they are more easily traced in polyphonic music by humans. Also, there is a preference for notes sounding in a middle or higher register compared to notes in the bass region.

## 3 EVALUATION

The aim of the MIREX Audio Melody Contest is to extract melodic content from polyphonic audio. The dataset contains 25 phrase excerpts of 10-40 seconds length from the following genres: Rock, R&B, Pop, Jazz, Solo classical piano. The ground truth data consists of a succession of frequency values with a spacing of 10 ms. Zero frequencies indicate periods without melody. The estimated frequency was considered correct whenever the corresponding ground truth frequency is within a range of 25 cents.

To maximise the number of possible submissions the melody transcription problem was divided into two subtasks, namely the melody pitch estimation and the distinction of melody and non-melody parts (voiced/unvoiced detection). Moreover it was possible to give a pitch estimate even for those parts, which have been declared unvoiced. Those frequencies are marked with a negative sign. As a consequence there are three possible algorithm output styles[2]:

- No segmentation (Brossier, Goto, Vincent & Plumbley)
- Segmentation without additional negative frequencies in detected unvoiced frames (Dressler, Paiva)
- Segmentation with additional negative frequencies in

[1] We do not identify individual notes literally, because there is no detection for hard and soft note onsets to split a stream into single notes. Nevertheless hard onsets and fast frequency changes may cause pauses in the sinusoidal trajectories, so that stream objects often correspond to notes. However we do not assign a discrete tone height and a discrete duration to the stream objects.

[2] A detailed description of the evaluation procedure and the results can be found online at http://www.music-ir.org/evaluation/mirex-results/audio-melody/index.html .

| Rank | Participant | Voicing Detection | Voicing False Alarm | Voicing d-prime | Raw Pitch Accuracy | Raw Chroma Accuracy | Overall Accuracy | Runtime (s) |
|---|---|---|---|---|---|---|---|---|
| 1 | Dressler | 81.8% | 17.3% | **1.85** | 68.1% | 71.4% | **71.4**% | 32 |
| **2** | Ryynänen & Klapuri | 90.3% | 39.5% | 1.56 | **68.6%** | **74.1%** | 64.3% | 10970 |
| **3** | Poliner & Ellis | 91.6% | 42.7% | 1.56 | 67.3% | 73.4% | 61.1% | 5471 |
| **3** | Paiva 2 | 68.8% | 23.2% | 1.22 | 58.5% | 62.0% | 61.1% | 45618 |
| **5** | Marolt | 72.7% | 32.4% | 1.06 | 60.1% | 67.1% | 59.5% | 12461 |
| **6** | Paiva 1 | 83.4% | 55.8% | 0.83 | 62.7% | 66.7% | 57.8% | 44312 |
| **7** | Goto | 99.9% * | 99.4% * | 0.59 * | 65.8% | 71.8% | 49.9% * | 211 |
| **8** | Vincent & Plumbley 1 | 96.1% * | 93.7% * | 0.23 * | 59.8% | 67.6% | 47.9% * | ? |
| **9** | Vincent & Plumbley 2 | 99.6% * | 96.4% * | 0.86 * | 59.6% | 71.1% | 46.4% * | 251 |
| **10** | Brossier † | 99.2% * | 98.8% * | 0.14 * | 3.9% | 8.1% | 3.2% * | 41 |

**Notes:** Bold numbers are the best in each column.
* Goto, Vincent and Brossier did not perform voiced/unvoiced detection, so the starred results cannot be meaningfully compared to other systems.
† Scores for Brossier are artificially low due to an unresolved algorithmic issue.

Table 1: 2005 MIREX Contest Results - Audio Melody Contest

detected unvoiced frames (Marolt, Poliner & Ellis, Ryännen & Klapuri)

### 3.1 Evaluation Results

The results of the contest are presented in Table 1. The statistics Voicing Detection, Voicing False Alarm and especially Voicing d-prime are indicators for the quality of the voiced/unvoiced detection, while Raw Pitch Accuracy and Raw Chroma Accuracy measure the quality of the melody pitch detection. The Overall Accuracy is the most important statistic, because it evaluates the segmentation as well as the pitch detection. Reaching 71.4% our submission has performed best on Overall Accuracy with a significant difference to other submissions (second best: 64.3%). As we have also reached the best result in the Voicing d-prime measure, we conclude that the good segmentation of voiced and unvoiced parts provides an important basis for this success.

Nevertheless we want to point out that even our pitch estimation approach is fully competitive. Since the calculation of the Raw Pitch Accuracy and the Raw Chroma Accuracy also include the negative frequencies in unvoiced frames, algorithms without this additional information are at a small disadvantage in comparison with algorithms, which provide additional negative frequencies, or algorithms without segmentation. If we only consider the correctly transcribed voiced instants our approach performs best, even though algorithms without segmentation should gain better results here, because they detect all voiced instants.

### 3.2 Performance Features

The algorithm is programmed in C++ and compiled as Windows™ binary. The performance of the algorithm varies slightly depending on the complexity of the audio input. The reported execution time for the MIREX 2005 test set, which consists of 25 audio pieces with an overall length of 685 seconds, is 32 seconds. Thus the time for the audio analysis is approximately 0.05 times real-time on an Intel® Pentium® 4 3.0 GHz CPU system with 3 GB RAM – the fastest runtime among all submissions.

Even though the MIREX runtime measures of the submissions cannot be compared directly, because the algorithms are implemented in different programming languages and have run on diverse computers and operating systems, significant differences can be observed. The execution times vary from 0.05 times real time (faster than real time) to about 65 times real time – more than three orders of magnitude. Our algorithm needed only 32 seconds while the second best algorithm needed 10970 seconds. We have already pointed out that we payed special attention to algorithm efficiency, but this would only explain minor differences.

## 4 CONCLUSION AND FUTURE WORK

In this paper we presented an overview of our submitted algorithm, which was named winner of the MIREX 2005 audio melody extraction contest. We showed that the use of efficient standard signal processing methods does not contradict with a good melody extraction performance. There still is a great potential for improvement through the application of knowledge from research areas related to music information retrieval (MIR). From the results of the MIREX evaluation we see, that a successful voiced/unvoiced detection is crucial for a decent melody transcription. Even though our approach performed best on this task, there is still room for refinement.

The proposed algorithm is presently limited to the extraction of a continuous melody contour. For the efficient application in MIR systems the continuous melody stream has to be segmented into a note-like representation. As part of our continuing research, we plan to address this challenging problem.

## ACKNOWLEDGEMENTS

Special thanks to the organisation team of the ISMIR 2004 Graduate School in Barcelona. Highlighting different as-

pects of music information retrieval this event has been a great source of references and inspiration.

Primary funding for this research was provided by *Landesgraduiertenförderung Thüringen*. The author also gratefully acknowledges support from the University of Konstanz providing assistance and facilities for this work.

## References

A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, Mass., 1st mit press paperback edition, 1994.

F. J. Charpentier. Pitch detection using the short-term phase spectrum. In *ICASSP*, pages 113–116, 1986.

J. L. Flanagan and R. M. Golden. Phase vocoder. *Bell System Technical Journal*, pages 1493–1509, 1966.

M. Goto. A robust predominant-f0 estimation method for real-time detection of melody and bass lines in cd recordings. In *ICASSP 2000, pp. II-757-760*, 2000.

M. Goto and S. Hayamizu. A real-time music scene description system: Detecting melody and bass lines in audio signals. *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis*, pages 31–40, 1999.

T. Ooura. General Purpose FFT (Fast Fourier/Cosine/Sine Transform) Package. http://momonga.t.u-tokyo.ac.jp/~ooura/fft.html, 2001.

X. Serra. *A System for Sound Analysis/ Transformation/ Synthesis based on a Deterministic plus Stochastic Decomposition*. PhD thesis, Stanford University, 1989.

B. Snyder. *Music and Memory: An Introduction*. MIT Press, Cambridge, Mass., 2000.