An Auditory Streaming Approach on Melody Extraction

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

Abstract

The MIREX (Music Information Retrieval Evaluation eXchange) framework provides a common set of data to evaluate and compare a vast variety of MIR systems. This paper describes our submission to the audio melody extraction evaluation addressing the task of identifying the melody pitch contour from polyphonic musical audio. It shall give an overview about the used methods and a discussion of the evaluation results. The presented algorithm is a derivative of our submission to MIREX'05. Therefor we will outline changes between the two versions and discuss the impact of the further developments.

The MIREX 2006 evaluation results show that our algorithm performs best in pitch detection and melody extraction.

Keywords: MIREX 2006, audio melody extraction.

1. METHOD

1.1. 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 a Multi Resolution FFT – an efficient technique used to compute STFT spectra in different time-frequency resolutions [1]. 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 [2] for the IF extraction.

1.2. Peak Selection

Sinusoidal components of the audio signal contain the most relevant information about the melody. Yet, it is a challenge to reliably identify sinusoidal partials in polyphonic music. Of course a consistent and moderate change in magnitude and frequency of the examined spectral peaks is a good criterium for the identification of sinusoidals. However, this requires a continuous tracking of partials with time, a demand which cannot be implemented easily for polyphonic audio signals.

Charpentier found that you can identify sinusoidals by distinct spectral features in one FFT frame alone [3]. We developed his method further and this way improved the performance and robustness of the adjacent pitch estimation noticeably. Nevertheless this efficient method is not adequate for audio signals with a dense spectrum, because it relies on a non distorted phase spectrum around spectral peaks. This is not the case for closely adjoining partials, which will be erroneously identified as noise and will be discarded from further analysis. For this reason we employed a psychoacoustic model in this year's application in contrast to the local sinusoidality criterion we applied in our submission to MIREX'05 [4].

Unlike the before-mentioned sinusoidality criterion, psychoacoustic masking is a method to exclude non audible peaks - sinusoidal or not - from further processing. We use a very simplified implementation of simultaneous and temporary masking, which by far does not reach the complexity of models used in modern lossy audio coders. However, this way many unprofitable peaks can be erased from the spectrum in order to speed up the further processing.

1.3. Pitch Estimation

The magnitude and instantaneous frequency of the sinusoids are evaluated by a pitch estimation method, as the frequency of the strongest harmonic may not be the perceived pitch of a periodic complex tone. At first, the pitch estimator performs a magnitude weighting and then it analyzes the harmonic structure of the polyphonic signal. The algorithm covers four octaves – computing pitch frequencies and an approximate prediction of the pitch salience in a frequency range between 80 Hz and 1280 Hz. A variable number of pitch candidates at each frame (about five pitches on average) is used to track tone objects.

1.4. Auditory Streaming

At the same time the frame-wise estimated pitch candidates are processed to build acoustic streams. Tones which have

a sufficient magnitude and are located in an adequate frequency range are assigned to the corresponding streams. Anyhow, every stream may possess only one active tone at any time. So in competitive situations the active tone is chosen with the help of a rating method that evaluates the tone magnitude and the frequency difference between the pitch of the tone and the actual stream position. Conversely, a tone is exclusively linked to only one stream.

This is a new concept compared to the method we used in last year. There, all tone objects lasting longer than 100 ms were grouped according to their frequency range and stored in different registers. Then all tone objects belonging the most energetic frequency range gained an additional weight in the concluding comparison. Yet, essentially any tone from any frequency region – even outside the most energetic frequency range – could win the final comparison and become part of the melody.

This is not the case in this year's algorithm where only tones from the most salient stream are considered to be melody tones. Therefor, the correct identification of the melody stream is very important for the success of the method!

1.5. Identification of the Melody Stream

Finally, the melody stream must be chosen. In general the most salient stream is identified as the melody. Of course it may happen that two ore more streams have about the same magnitude and thus no clear decision can be taken. In this case, the stream magnitudes are weighted according to their frequency. Streams from the bass region receive a lower weight than streams from the mid and high frequency regions. If no clear melody stream emerges during a short time span, the most salient weighted stream is chosen.

2. Implementation

The algorithm is implemented in C++ and is available for Windows and Linux platforms. The performance of the algorithm varies slightly depending on the complexity of the audio input. The reported execution time for the MIREX 2006 test sets, which consist of 45 audio pieces with an overall length of 1053 seconds, is 75 seconds. Thus the audio analysis is approximately 14 times faster than real-time on an AMD Athlon XP 2600+1.9GHz CPU system with 2 GB RAM – the fastest runtime among all submissions. The implementation is suitable for the instant processing of an audio stream, although with a latency of 250ms up to 4s this implementation is not suitable for real-time processing. However, the allowed latency may be decreased to a minimum value of about 25ms. Of course such a small latency will noticeably decrease the overall accuracy of the algorithm.

3. MIREX Evaluation

3.1. Evaluation Overview

The aim of the MIREX Audio Melody Evaluation is to extract melodic content from polyphonic audio. Two datasets were available for the evaluation this year. The MIREX 2005 dataset contains 25 phrase excerpts of 10-40 seconds length from the following genres: Rock, R&B, Pop, Jazz, Solo classical piano. The same data was used for the MIREX 2005 audio melody contest. This way a direct comparison between the evaluation 2005 and this year's evaluation is possible. For the ISMIR 2004 Audio Description Contest, the Music Technology Group of the Pompeu Fabra University assembled a diverse set of 20 polyphonic musical audio pieces and corresponding melody transcriptions including MIDI, Jazz, Pop and Opera music as well as audio pieces with a synthesized voice. Each file has an approximate duration of 20 seconds [1] .

The audio excerpts are provided as single channel PCM data in CD-quality (16-bit resolution, 44.1 kHz sample rate). The corresponding reference annotations of the predominant melody include a succession of pitch frequency estimates at discrete time instants (5.8/10 ms grid). Zero frequencies indicate periods without melody. The estimated frequency was considered correct whenever the corresponding ground truth frequency is within a range of 50 cents.

To maximise the number of possible submissions the transcription problem was divided into two subtasks, namely the melody pitch estimation and the distinction of melody and non-melody parts (voiced/unvoiced detection). 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. Moreover, each dataset was divided into a vocal and a non-vocal melody voice subset.

3.2. Results

The evaluation results show that our algorithm performs best in pitch detection and melody extraction [2] . As indicated by the excellent runtime of our algorithm, the implemented methods allow a very efficient computation of the melody pitch contour.

Table 1 shows that the overall accuracy varies significantly among the submissions. However, we must not forget that quite different approaches are compared. In contrast to the other transcription systems, Poliner and Ellis present a classification-based system that uses no assumptions about the physical nature of sound [5]. Brossier aims at real-time processing with a very short latency [6]. Sutton et al have built a system that is only suitable for singing voice extraction [7]. So naturally their system performs better for the vocal pieces. Ryynänen and Klapuri use a general approach with a parameter setting especially tuned for the transcription of the singing voice [8]. For the given vocal examples,

[1] The data set including the reference annotations can be found on the contest web page http://ismir2004.ismir.net/melody_contest/results.html

[2] Detailed evaluation results can be found at http://www.music-ir.org/mirex2006/index.php/Audio_Melody_Extraction_Results

Table 1. 2006 MIREX Audio Melody Extraction Results

Dataset	Participant	Voicing Recall	Voicing False Alm	Voicing d-prime	Raw Pitch	Raw Chroma	Overall Accuracy	Runtime (s)
ISMIR 2004	Dressler	90.9%	10.5%	2.58	82.9%	84.0%	82.5%	27
	Ryynänen & Klapuri	84.4%	12.6%	2.16	80.6%	82.3%	77.3%	440
	Poliner & Ellis	89.9%	36.3%	1.63	73.2%	76.4%	71.9%	-
	Sutton et al	73.2%	24.9%	1.30	62.6%	65.4%	58.2%	5014
	Brossier	99.7%	88.4%	1.61	57.4%	68.7%	49.6% *	30
MIREX 2005	Dressler	89.3%	28.8%	1.80	77.7%	82.0%	73.2%	48
	Dressler (2005)	*81.8%*	*17.3%*	*1.85*	*68.1%*	*71.4%*	*71.4%*	*32*
	Ryynänen & Klapuri	78.2%	16.5%	1.75	71.5%	75.0%	67.9%	773
	Poliner & Ellis	93.5%	45.1%	1.64	66.2%	70.4%	63.0%	-
	Sutton et al	64.5%	13.8%	1.46	56.4%	60.1%	53.7%	8195
	Brossier	99.5%	98.2%	0.46	41.0%	56.1%	31.9% *	58

Note: * Brossier did not perform voiced/unvoiced detection, so the overall accuracy cannot be meaningfully compared to other systems.

no significant difference can be noted between the accuracy of this implementation and our submission.

All resubmitted algorithms have improved the overall accuracy compared to the results of MIREX'05. As we can see in table 1 (where the submission of last year is marked by italic font), our melody extraction algorithm has gained 1.8% in overall accuracy for the MIREX 2005 dataset. The improvements for the raw pitch and raw chroma estimation seem even more pronounced. Yet, a part of this increased accuracy has to be attributed to the use of the negative frequency output, which has not been used last year.

4. Acknowledgements

Many thanks to the IMIRSEL team at the at the University of Illinois at Urbana-Champaign for running the MIREX evaluations.

This research was partially founded by the state of Thuringia, Germany (*Landesgraduiertenförderung Thüringen*).

References

[1] K. Dressler, "Sinusoidal extraction using an efficient implementation of a multi-resolution FFT," in *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-06)*, Montreal, Quebec, Canada, Sept. 18–20, 2006, pp. 247–252.

[2] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell System Technical Journal*, pp. 1493–1509, 1966.

[3] F. J. Charpentier, "Pitch detection using the short-term phase spectrum," in *Proc. of ICASSP 86*, 1986, pp. 113–116.

[4] K. Dressler, "Extraction of the melody pitch contour from polyphonic audio," in *1st Music Information Retrieval Evaluation eXchange (MIREX)*, 2005, available online http://www.musicir.org/evaluation/mirex-results/articles/melody/dressler.pdf.

[5] G. E. Poliner and D. P. W. Ellis, "Classification-based melody transcription," *Machine Learning Journal*, 2006.

[6] P. Brossier, J. P. Bello and M. D. Plumbley, "Real-time temporal segmentation of note objects in music signals," in *Proceedings of the International Computer Music Conference (ICMC 2004)*, Florida, USA, 2004.

[7] C. Sutton, E. Vincent, M. D. Plumbley and J. P. Bello, "Transcription of vocal melodies using voice characteristics and algorithm fusion," in *2nd Music Information Retrieval Evaluation eXchange (MIREX)*, 2006, available online http://www.music-ir.org/evaluation/MIREX/2006_abstracts/AME_sutton.pdf.

[8] M. P. Ryynnen and A. P. Klapuri, "Transcription of the singing melody in polyphonic music," in *Proc. 7th International Conference on Music Information Retrieval (ISMIR)*, Victoria, Canada, 2006.