AUDIO MELODY EXTRACTION FOR MIREX 2009

Karin Dressler
Fraunhofer IDMT, Ilmenau, Germany
kadressler@gmail.com

ABSTRACT

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 algorithm and a discussion of the evaluation results. The presented algorithm is a derivative of our submission to MIREX'06. Major changes between the two versions are highlighted and the impact of the further developments is discussed.

The MIREX 2009 evaluation results show that the presented algorithm has the best overall accuracy in melody pitch 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 amounts of zero padding using a Hann window. Thereby a Multi Resolution FFT is used – 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.

1.2 Peak Selection

The more elaborate peak selection method of the last algorithm version [4] has been replaced by a simple magnitude threshold. The threshold depends on the signal, as it is a fraction of the biggest peak magnitude of the current frame. Actually, the fraction is chosen to be very small so that almost all peaks will be processed further.

Then the instantaneous frequency (IF) for the selected peaks is computed. In order to obtain more stable IF measures, the average of two estimation methods is used, namely the well-known phase vocoder [2] and a method proposed by Charpentier [3].

1.3 Pitch Estimation

The weighted magnitude and the instantaneous frequency of the selected spectral peaks are evaluated in order to identify the strongest periodicities. The periodicity estimation is based on the pair-wise analysis of spectral peaks. The main idea lies in the identification of partials with a successive harmonic number. Consecutively, the identified harmonic peak pairs are evaluated according to a perceptually motivated rating scheme. The resulting pitch strengths are then added to a pitch spectrogram. Pitch frequencies and an approximate prediction of the pitch salience are computed in a frequency range between 55 Hz and 1318 Hz.

1.4 Tones

In contrast to our last MIREX submission (where the most salient pitches in each frame were used to track tone objects) the salient pitches function only as starting points for new tone objects. The actual estimation of tone height and tone magnitude is performed as an independent computation: harmonic peaks are added to existing tone objects and after a short time (e.g. a few frames) a timbre representation for that tone is established. The timbre determines how much harmonic partials of the current frame will influence pitch and magnitude of the tone. This way the impact of noise and other sound sources can be decreased noticeably.

1.5 Auditory Streaming

At the same time the frame-wise estimated pitch candidates are processed to build acoustic streams. A rating is calculated for each tone depending on loudness, frequency dynamics, tone salience and tone to voice distance. Tones with a sufficient rating 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 tone height and the actual stream position. Conversely, a tone is exclusively linked to only one stream.

1.6 Identification of the Melody Stream

Finally, the melody voice must be chosen. In general the most salient auditory 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

Participant	Vx Recall(%)	Vx False Alm(%)	Raw Pitch(%)	Raw Chroma(%)	Overall Acc(%)	Runtime (min)
cl1	93.01	80.71	63.45	66.29	52.19	28
cl2	80.30	57.42	63.45	66.29	55.19	33
dr1	92.40	51.74	74.45	76.82	66.86	23040
dr2	87.69	41.22	72.09	75.72	66.17	524
hjc1	43.62	9.71	66.12	72.58	50.49	344
hjc2	43.62	9.71	51.13	67.12	49.01	584
jjy	61.02	29.39	73.33	79.68	56.64	3726
kd	90.93	40.99	80.58	82.52	73.35	24
pc	79.32	40.29	64.10	65.84	62.88	4677
rr	91.28	51.11	72.21	76.33	65.22	26
toos*	99.87	98.31	75.05	80.34	55.08	1468
mw*	99.97	98.66	73.44	77.50	55.07	132

Table 1. MIREX 2009 Audio Melody Extraction Overall Summary Results - Unweighted Avg. of all Datasets

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. MIREX EVALUATION

2.1 Evaluation Overview

The aim of the MIREX Audio Melody Evaluation is to extract melodic content from polyphonic audio. Four datasets were available for the evaluation this year.

- MIREX09: 374 excerpts of 20-40s of Chinese Karaoke songs (singing voice, synthetic accompaniment). The same database was tested with different melodic voice to accompaniment ratios. (+5dB, 0dB, and -5 dB RMS)
- MIREX08: 8 excerpts of 60s from north Indian classical vocal performances.
- MIREX05: 25 excerpts of 10-40s from the following genres: Rock, R&B, Pop, Jazz, Solo classical piano.
- ADC04: 20 excerpts of about 20s including MIDI, Jazz, Pop and Opera music as well as audio pieces with a synthesized voice.

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 time periods without melody. The estimated frequency was considered correct whenever the corresponding ground truth frequency is within a range of 100 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 an additional pitch estimate for the frames that are declared unvoiced by the algorithm. Those frequencies are marked with a negative sign. Moreover, each dataset was divided into a vocal and a non-vocal melody voice subset.

2.2 Evaluation Results

The MIREX 2009 evaluation results show that the presented algorithm has the best Overall Accuracy in melody pitch extraction regardless of the tested dataset. Unlike some other algorithms, there is no performance break-down for audio input with an instrumental lead voice, as the proposed method is not specifically adapted to a human melody voice.

The unweighted average of the evaluation results for all datasets is shown in Table 1, where our submission is highlighted in grey. The Overall Accuracy is the most important statistic, because it evaluates the segmentation between melody and non-melody parts as well as the pitch detection [1] . Reaching 73.3% our submission has performed best on Overall Accuracy with a significant difference to other submissions (second best: 66.9%).

As indicated by the excellent runtime of our algorithm, the implemented methods allow a very efficient computation of the melody pitch contour. The submitted method runs clearly faster than the second-ranked algorithm (960 times faster), although it should be noted that run times cannot be compared directly because the algorithms are implemented in different programming languages and have run on diverse computers. The best system with a comparable run time reached 65.2% Overall Accuracy.

Figure 1 shows the trend of the MIREX audio melody evaluation results for different data sets [2] . There is clearly much room for improvement in the future, though it can be noted that there was no bold increase in the melody extraction accuracy during the last three years. In current systems, melody and accompaniment are mainly separated by their sound intensity. It is obvious that much better results cannot be achieved with this feature alone. For example, the human voice has a high dynamic range (about 20 dB) and the instrumental accompaniment naturally reaches the volume of the softer human melody parts. Thus more high level features have to be incorporated into the melody

[1] The starred submissions did not perform voiced/unvoiced detection, so the overall accuracy cannot be meaningfully compared to other systems.

[2] As the data set ADC 04 was no official test set in the year 2005, the graph shows the performance of the author's submission to MIREX 2005 on this data set.

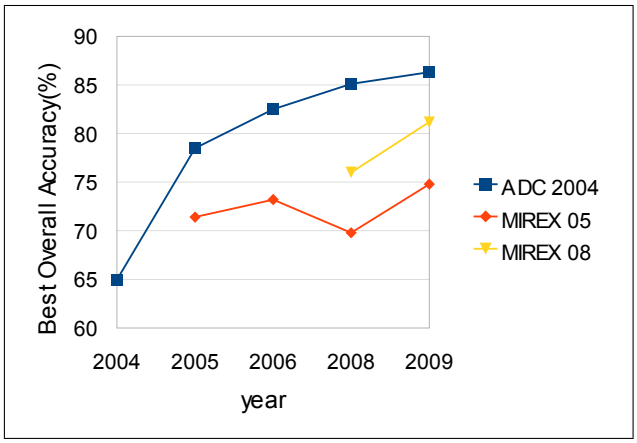

Figure 1. Evaluation Trends in Audio Melody Extraction

extraction process, such as the loudness envelope, the frequency evolution and the timbre of a tone.

3. 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, "An Auditory Streaming Approach on Melody Extraction," in *2nd Music Information Retrieval Evaluation eXchange (MIREX)*, 2006, available online http://www.music-ir.org/evaluation /MIREX/2006_abstracts/AME_dressler.pdf.