AUDIO MELODY EXTRACTION FOR MIREX 2014

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 MIREX 2014 evaluation results show that the presented algorithm has the best overall accuracy in melody pitch extraction among the participating algorithms.

1. METHOD

Two algorithms have been submitted to the MIREX audio melody extraction task this year. The algorithm KD3 is basically our submission of the year 2009 [2], the algorithm KD1 incorporates the tone processing that was developed for the MIREX multiple fundamental frequency estimation task.

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The more elaborate peak selection described in [1] has been replaced by a simple magnitude threshold.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 (maximum magnitude multiplied with 0.01) 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 [3] and a method proposed by Charpentier [4].

1.3 Pitch Estimation

The weighted magnitude and the instantaneous frequency of the selected spectral peaks are evaluated in order to identify the predominant fundamental frequency. The periodicity estimation is based on the pair-wise analysis of spectral peaks [5]. 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. The salient pitches are then connected to form short pitch tracks, which help to identify starting points for new tones.

1.4 Tones

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 spectral envelope for that tone is established, which evaluates smoothness constraints for harmonics over time and frequency. Furthermore, the problem of shared harmonics is addressed. The long term spectral envelope determines how much a harmonic partial 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 updated tones are processed to build acoustic streams [6]. 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.

Algorithm	Overall Accuracy (%)	Raw Pitch (%)	Raw Chroma (%)	Voicing Recall (%)	Voicing False Alarm(%)
SG2 (2011)	**75.1**	79.5	82.3	86	23.7
KD3	73.3	**80.6**	**82.5**	90.9	41.0
KD1	72.2	79.3	81.3	86.4	32.5
DD1	71.4	72.2	75.0	85.9	29.6
CWJ3	67.5	73.4	75.1	73.8	19.7
IYI1	59.8	68.9	73.1	84.5	39.3
SL1	57.1	52.4	55.3	73.2	22.6
YJ2*	54.0	74.8	78.6	100.0	99.6
LPSL1	40.6	40.5	46.6	67.8	35.9

Table 1. MIREX 2014 Audio Melody Extraction Overall Summary Results - Unweighted Avg. of all Datasets

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 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 energy 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 unweighted average of the evaluation results for all datasets is shown in Table 1, where our submissions are denoted by the submission shortcodes KD1 and KD3. The MIREX 2014 evaluation results show that the presented algorithms have the best Overall Accuracy for the unweighted average of all datasets among the participating algorithms. However, they do not reach the outstanding result of Justin Salamon's submission for MIREX 2011 [7].

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] . It seems that our algorithms cope well with instrumental music, as there is no performance break-down for audio input with an instrumental lead voice. One reason (besides the presented approach to auditory streaming) might be that the proposed method is not specifically adapted to a human melody voice. This might also explain the moderate results for the MIREX09 database which contains Chinese Karaoke songs.

Unfortunately, the new multiple F0 estimation front end, which was developed for the MIREX multiple F0 estimation and tracking task, does not improve the overall accuracy of the melody extraction. A reason might be that the multiple F0 estimation enhances the tones of the accompaniment, so that it becomes more difficult to extract the predominant melody voice.

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

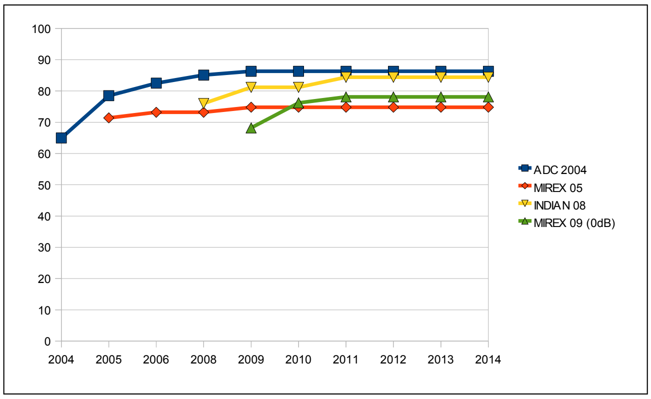

Figure 1. Improvement of Overall Accuracy in Audio Melody Extraction

Figure 1 shows the best 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 increase in the melody extraction accuracy since 2011. It seems as if the algorithm performance has reached a glass ceiling that is not easy to overcome. 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 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] K. Dressler, “Audio Melody Extraction for MIREX 2009,” *5th Music Information Retrieval Evaluation eXchange (MIREX)*, 2009.

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

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

[5] K. Dressler, “Pitch estimation by the pair-wise evaluation of spectral peaks,” in *Proc. AES 42nd International Conference*, Ilmenau, Germany, 2011.

[6] K. Dressler, “Towards Computational Auditory Scene Analysis: Melody Extraction from Polyphonic Music,” in *Proc. of 9th Int. Symposium on Computer Music Modelling and Retrieval (CMMR 2012)*, London, UK, 2009.

[7] J. Salamon and E. Gómez, “Melody extraction from polyphonic music: MIREX 2011,” *7th Music Information Retrieval Evaluation eXchange (MIREX)*, 2011.

[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.