AUDIO MELODY EXTRACTION FOR MIREX 2020

Karin Dressler
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.

1. METHOD

1.1 Spectral Analysis and Magnitude Weighting

If a partial of a complex tone is not obscured by other harmonics or noise, it can be detected as a peak in the magnitude spectrum of the Short Term Fourier Transform (STFT). The interference of partials from simultaneously playing notes can be decreased if the frequency resolution of the STFT is increased. However, musical sound is not stationary, so very long STFT data windows cannot be used to gain a very high frequency resolution. As a compromise between a good frequency resolution and a good time resolution, we analyze the audio signal by calculating a multi-resolution Fast Fourier Transform (MR FFT) [1].

The best frequency resolution ($\Delta f = 21.5$ Hz) is reached for the low frequency components up to approximately 600 Hz. The best time resolution corresponds to a FFT data window length of 5.8 ms for frequencies above 4400 Hz. Due to different amounts of zero padding the resulting STFT frame length and the hop size of the analysis window correspond to 2048 samples and 5.8 ms for all STFT resolutions (for music sampled at 44.1 kHz).

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].

In order to obtain the weighted magnitude for the spectral peak , its STFT magnitude is multiplied with the peak's instantaneous frequency. This weighting introduces a 6 dB magnitude boost per octave. In effect, the weighted signal is proportional to the signal derivative.

1.2 Pitch Estimation

For the computation of the pitch spectrogram, spectral peaks in the frequency range between 55 Hz and 5 kHz are processed. The weighted magnitude and the instantaneous frequency of the spectral peaks are evaluated in order to identify the strongest signal periodicity in the frequency range between 55 Hz and 2093 Hz. The pitch estimation algorithm is based on the pair-wise analysis of spectral peaks [4]. The idea of the technique lies in the identification of partials with successive (odd) harmonic numbers. Since successive partials of a harmonic sound have well defined frequency ratios, a possible fundamental frequency (F0) can be derived from the instantaneous frequencies of two spectral peaks. Consecutively, the identified harmonic pairs are rated according to timbral smoothness and common frequency deviation. Finally, the resulting pitch strengths are added to a pitch spectrogram. Then, short pitch tracks are build from salient pitches in order to identify the predominant periodicity.

1.3 Tones

The tone estimation allows the parallel tracking of up to ten tones. The novell approach is described in detail in [5]. A high level tone object is started from a pitch track, if the best rated one passes an adaptive magnitude threshold.

All active tone objects are jointly evaluated over time in order to estimate their pitch and their magnitude. At the same time a spectral envelope is established for each tone. The spectral envelope (e.g. harmonic magnitudes) determines the weight each spectral peak receives in the tone's pitch and magnitude estimation. In this way, the impact of noise and concurrent tones can be decreased noticeably.

In order to establish long term timbre information, adequate spectral peaks are assigned to the active tone objects in each analysis frame. The added spectral peaks, eventual masking and the computed tone height are exploited in a rating scheme that determines how well each harmonic can be integrated into the overall timbre. A feedback about the existing tone objects is provided to the pitch determination method, so that matched spectral peaks can be inhibited during the pitch determination. This way, pitches besides the predominant one can be extracted.

1.4 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.

1.5 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. EVALUATION

2.1 Evaluation Metrics

2.2 Results and Discussion

3. CONCLUSION

4. 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, "Pitch estimation by the pair-wise evaluation of spectral peaks," in *Proc. AES 42nd International Conference*, Ilmenau, Germany, 2011.

[5] K. Dressler, "Automatic transcription of the melody from polyphonic music," *Ph.D. thesis*, Ilmenau, Germany, 2017.

[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.