MULTIPLE FUNDAMENTAL FREQUENCY EXTRACTION FOR MIREX 2011

Karin Dressler
Fraunhofer Institute for Digital Media Technology IDMT, Ilmenau, Germany
kadressler@gmail.com

ABSTRACT

This extended abstract outlines an efficient approach for the extraction of multiple fundamental frequencies from polyphonic musical audio. The algorithm consists of three analysis steps. At first a multi resolution spectral analysis is performed on the audio signal. Then, the most salient pitches are identified using a pitch extraction algorithm, which is designed to identify the predominant pitch in polyphonic audio. Finally, distinct high level tone objects are created and tracked over time: the most salient pitch of the current analysis frame may start a new tone object. All active tone objects are jointly evaluated in order to estimate their pitch and magnitude and establish timbre information.

The proposed algorithm is a front-end for a melody extraction system, which places a high priority at the most salient tones and at the processing of a human singing voice. Hence, the evaluation results for the multiple-F0 analysis are also discussed in the scope of melody extraction.

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 size and the hop size of the analysis window are 2048 and 256 samples for all STFT resolutions, respectively. Assuming audio data sampled at 44.1 kHz, the highest frequency resolution corresponds to an FFT window length of 46 ms, and the hopsize is 5.8 ms.

In order to obtain the weighted magnitude A_s for the spectral peak at STFT bin k, its STFT magnitude is multiplied with the peak's instantaneous frequency f_i.

$$A_s[k] = |X[k]| \cdot f_i[k] \tag{1}$$

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

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. For the computation of the pitch spectrogram, spectral peaks in the frequency range between 55 Hz and 5 kHz are processed. The pitch estimation algorithm is based on the pair-wise analysis of spectral peaks [2]. 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 the two spectral peaks. Consecutively, the identified harmonic pairs are rated according to harmonicity, timbral smoothness, the appearance of intermediate spectral peaks and harmonic number. Finally, the resulting pitch strengths are added to a pitch spectrogram.

1.3 Tones

A high level tone object is started, if the most salient pitch in the current analysis frame 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. The principle indicators for the harmonic fit are: 1) the frequency difference between tone height and computed virtual pitch of the harmonic, 2) the smoothness of the timbre in the frequency and time dimension, and 3) the magnitude division of shared harmonics among distinct tones.

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 pitch can be extracted.

2. EVALUATION

The presented method for the detection of multiple F0 has been implemented as part of a melody extraction algorithm, which was evaluated at the Music Information Retrieval Evaluation eXchange (MIREX) in 2009 [3]. Algorithm parameters (for example the FFT window size, parameters for the pitch estimation and tone tracking, as well as the timing constants of the adaptive thresholds) have been adjusted using the melody extraction training data of ISMIR 2004 and MIREX 2005. However, there is one modification for the MIREX multiple-F0 task: the maximum allowed frequency range for tones was increased to cover frequencies between 55 Hz and 2093 Hz. Nonetheless, it should be noted that the used parameter setting is probably not the best choice to maximize the estimation accuracy for the multiple-F0 task – in particular, as the dataset for melody extraction consists mostly of musical pieces with a singing voice, while the multiple-F0 dataset includes solely instrumental music. So usually, most multiple-F0 estimation algorithms use a longer FFT frame length, because the frequency of the instrumental music is relatively stable. Since our algorithm functions as a front end to a melody extraction algorithm, the priority is of course to detect the melody line, which often comprises a human singing voice.

2.1 Evaluation Metrics

Two different sets of evaluation metrics are used to estimate the algorithm performance in the multiple fundamental frequency estimation task. The first set estimates the algorithm performance in terms of precision, recall and overall accuracy using the following equations:

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

$$Accuracy = \frac{TP}{TP + FP + FN}, \tag{4}$$

where TP is the number of correctly identified pitches (true positives), FP is the number of identified pitches which do not occur in the ground truth (false positives), and FN is the number of pitches which are not identified by the algorithm (false negatives).

The second set of evaluation metrics was proposed by Poliner and Ellis in order to measure the accuracy of polyphonic piano transcriptions [4]. The metric computes an error score E_{tot}, that takes into account the so-called substitution errors E_{subs}, which allows the substitution of any false positive F0 with a ground-truth F0 which was not reported [4]. Moreover, the number of errors is set into relation to the total quantity of notes:

$$E_{\text{subs}} = \frac{\sum_{t=1}^{T} \min(N_{\text{ref}}(t), N_{\text{sys}}(t)) - N_{\text{corr}}(t)}{\sum_{t=1}^{T} N_{\text{ref}}(t)}, \tag{5}$$

where N_{ref} is the number of pitches in the ground truth data, N_{sys} is the number of pitches returned by the system, N_{corr} is the number of correctly identified pitches, and t is the index of the current analysis frame.

The remaining components of the metric are missing pitches E_{miss} and false alarm errors E_{fa}. While E_{miss} refers to the number of ground-truth reference notes that could not be matched with any system outputs (i.e. misses after substitutions are accounted for), E_{fa} refers to the number of pitches that cannot be paired with any ground truth (false alarms beyond substitutions):

$$E_{\text{miss}} = \frac{\sum_{t=1}^{T} \max(0, N_{\text{ref}}(t) - N_{\text{sys}}(t))}{\sum_{t=1}^{T} N_{\text{ref}}(t)} \tag{6}$$

$$E_{\text{fa}} = \frac{\sum_{t=1}^{T} \max(0, N_{\text{sys}}(t) - N_{\text{ref}}(t))}{\sum_{t=1}^{T} N_{\text{ref}}(t)}. \tag{7}$$

The total error is estimated as follows:

$$E_{\text{tot}} = \frac{\sum_{t=1}^{T} \max(N_{\text{ref}}(t), N_{\text{sys}}(t)) - N_{\text{corr}}(t)}{\sum_{t=1}^{T} N_{\text{ref}}(t)}. \tag{8}$$

	Precision (%)	Recall (%)	Accuracy (%)	Etot	Esubs	Emiss	Efa	Runtime (sec)
BD1	0.64	0.68	0.57	0.53	0.20	0.11	0.21	99135
KD1	0.85	0.66	0.63	**0.38**	0.08	0.26	0.04	149
LYC1	0.56	0.59	0.47	0.71	0.24	0.16	0.30	28138
RFF1	0.63	0.57	0.49	0.59	0.19	0.24	0.16	148477
RFF2	0.57	0.60	0.49	0.66	0.22	0.18	0.26	384676
YR1	0.73	0.80	0.66	0.43	0.09	0.11	0.23	6584
YR2	0.73	0.84	**0.68**	0.42	0.08	0.08	0.26	6584
YR3	0.71	0.80	0.65	0.46	0.10	0.10	0.26	6578
YR4	0.72	0.84	**0.68**	0.43	0.08	0.07	0.27	6578

Table 1. Multiple Fundamental Frequency Estimation Results of MIREX 2011

2.2 Results

Table 1 shows the analysis results for the MIREX multiple frequency estimation task. While the algorithm does not reach the overall accuracy of the systems submitted by Yeh and Roebel [5], our algorithm performs best in terms of the total error metric E_{fa} introduced by Polliner and Ellis in [4]. Moreover, the MIREX results show that the implemented algorithm obtains the highest Precision, e.g. 85% of the extracted fundamental frequencies are true positives. On the other hand it becomes obvious that the algorithm systematically underestimates the number of concurrent voices, leading to a low Recall of 66%.

However, this fact might not be very surprising, as the main purpose of a melody extraction algorithm is to extract the strongest notes of a musical piece – a full transcription of all notes is not necessary. While a better trade-off between Precision and Recall might be reached by using a more suitable dataset for the parameter estimation, it may not be possible to achieve a much better accuracy without loosing some generality in terms of the input data.

It can also be noted that the submitted algorithm stands out due to very short run-times.

3. CONCLUSION

In this extended abstract we presented an efficient approach to the estimation of multiple fundamental frequencies from polyphonic music. The MIREX results show that the proposed method allows a reliable and very efficient identification of the multiple fundamental frequencies.

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)*, pages 247–252, Montreal, Quebec, Canada, Sept. 2006.

[2] K. Dressler. Pitch estimation by the pair-wise evaluation of spectral peaks. In *AES 42nd Conference*, Ilmenau, Germany, July 2011.

[3] K. Dressler. Audio Melody Extraction for MIREX 2009. In *5th Music Information Retrieval Evaluation eXchange (MIREX)*, 2009.

[4] G.E. Poliner and D.P.W. Ellis. A discriminative model for polyphonic piano transcription. In *EURASIP Journal on Advances in Signal Processing*, 2007

[5] C. Yeh and A. Roebel Multiple-F0 estimation for MIREX 2011. In *7th Music Information Retrieval Evaluation eXchange (MIREX)*, 2011.