MELODY EXTRACTION FROM POLYPHONIC AUDIO SIGNAL MIREX2011

Sanghun Park **Seokhwan Jo** **Chang D. Yoo**
Dept. of EE
Korea Advanced Institute of Science and Technology
373-1 Guseong Dong, Yuseong Gu, Daejeon 305-701, Korea
psh111@kaist.ac.kr antiland@kaist.ac.kr cdyoo@ee.kaist.ac.kr

ABSTRACT

This paper considers the proposed algorithm submitted to the Music Information Retrieval Evaluation eXchange(MIREX) 2011 "Audio Melody Extraction" task. The proposed melody pitch extraction algorithm can be divided into three steps: (1) a spectral analysis using variable length window, (2) a pitch candidate estimation, and (3) a pitch sequence identification. In the first step, the short-time Fourier transform (STFT) with variable length window is performed to be robust against dynamic variation of melody line. In the second step, melody pitch candidates of each frame are obtained from weights of a harmonic structure in the spectrum. In the third step, a single pitch sequence (melody line) is selected from the many possible pitch sequences based on the general properties of melody line.

1. INTRODUCTION

The Music Information Retrieval Evaluation eXchange(MIREX) audio melody extraction contest has had considerable impact on the tremendous progress in the melody extraction over last decade. In spite of progress in the melody extraction [1–3], it is still difficult to improve an accuracy of melody extraction due to the following reasons: harmonic interference and octave mismatch [4].

In this competition, we propose a simple and effective melody extraction algorithm which is robust to the aforementioned difficulties. The proposed algorithm extracts the melody line in three steps: (1) a spectral analysis using variable length window, (2) a pitch candidate estimation, and (3) pitch sequence identification. In the first step, a transient analysis is performed on the polyphonic input audio to find a suitable window length of each frame. Then, the short-time Fourier transform (STFT) with variable length window is performed. In the second step, melody pitch candidates of each frame are obtained from weights of a harmonic structure in the spectrum. The effect of harmonic interference and octave mismatch can be reduced by considering several melody pitch candidates in each frame. In the third step, a single pitch sequence is identified from the melody pitch candidates based on a rule-based method. The overall structure of the proposed algorithm is shown in Figure 1.

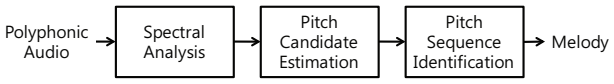


Figure 1. System Overview.

2. METHOD DESCRIPTION

2.1 Spectral analysis

To analyze the given polyphonic audio, the short-time Fourier transform(STFT) with variable length window is performed. The proposed algorithm uses a short window for transition regions and uses a long window for monotonous melody pitch regions in the range between 32ms and 92ms. The range of the window length is decided based on experimental results.

2.2 Pitch Candidate Estimation

In each frame, several melody pitch candidates are obtained to reduce the estimation errors due to harmonic interference and octave mismatch. To extract pitch candidates from polyphonic audio, the weights of the modified harmonic structure model proposed in [2] are estimated. The harmonic structure model is represented as

$$H_\omega(k) = \sum_{m=1}^{H} A_m G(k; \omega + 1200 \log_2 m, W), \qquad (1)$$

where ω, A_m, H and W are the fundamental frequency $F0$, the amplitude of mth harmonic partial, number of harmonics, and the variance of function G, respectively. Here, $G(x; x_0, \varsigma)$ is a Gaussian function defined as

$$G(x; x_0, \varsigma) = \frac{1}{\sqrt{2\pi\varsigma^2}} \exp\left[-\frac{(x - x_0)^2}{2\varsigma^2}\right]. \qquad (2)$$

Figure 2 illustrates the harmonic structure model used in the considered algorithm.

The weights are calculated as the inner-dot product between the harmonic structure with fundamental frequency

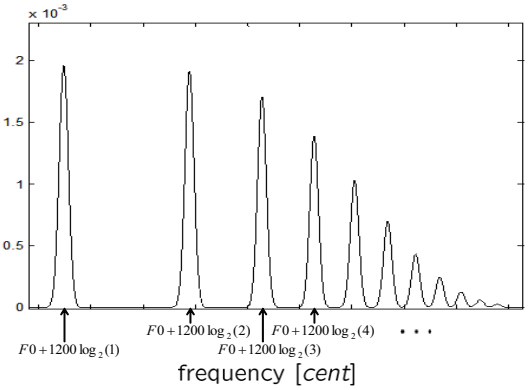


Figure 2. Harmonic structure model $H_\omega(k)$ of which $H = 11$.

ω and the spectral magnitudes. The weights of fundamental frequency ω in the lth frame is mathematically expressed as

$$J(\omega, l) = \sum_k |S(k, l)| H_\omega(k). \quad (3)$$

Here, $J(\omega, l)$ informs the strength of the harmonic structure of a pitch frequency ω in the lth frame. The pitch candidates are extracted by picking the peak values of $J(\omega, l)$. Figure 3 (a) illustrates a certain STFT magnitude, and Figure 3 (b) illustrates its $J(\omega, l)$. The circles in Figure 3 (b) indicate the melody pitch candidates.

2.3 Pitch Sequence Identification

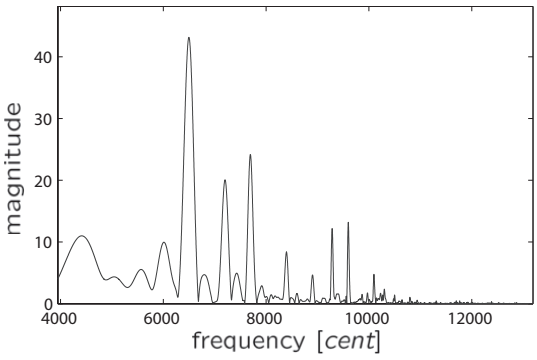
A simple way to estimate a melody pitch of each frame is to find the ω that maximizes the $J(\omega, l)$:

$$F_0(l) = \arg\max_\omega J(\omega, l). \quad (4)$$

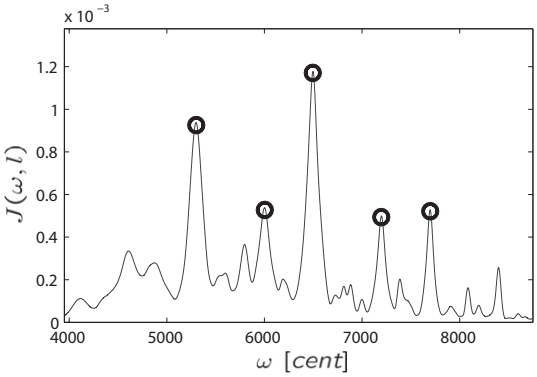
However, this result is not always reliable because of harmonic interference and octave mismatch. Therefore, we consider several pitch candidates to get more reliable result.

Posterior to the pitch candidate estimation step, L-best melody lines are estimated from N-best pitch candidates based on rule-based method. These rules are defined based on basic properties of the melody line [1, 5].

Once the pitch identification process is performed, any spurious pitch estimates are removed and replaced with a value interpolated between non-spurious estimates. There are spurious estimates after the identification process for the following two reasons: (1) N-best candidates may not include ground-truth pitch value, and (2) the rule discussed above is not complete to cover all possible situations.



(a) $S(k, l)$



(b) $J(\omega, l)$

Figure 3. $S(k, l)$ and its $J(\omega, l)$. The circles (o) indicate the melody pitch candidates.

3. REFERENCES

[1] G. E. Poliner, D. P. W. Ellis, and A. F. Ehmann: "Melody Transcription from Music Audio: Approach and Evaluation," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, N0. 4, pp. 1247–1256, 2007.

[2] M. Goto: "A real–time music–scene–description system: predominant–F0 estimation for detecting melody and bass lines in real–world audio signals," *Speech Communication*, Vol. 43, No. 4, pp. 311–329, 2004.

[3] R. P. Paiva, T. Mendes, and A. Cardoso: "Melody Detection in Polyphonic Musical Siganls: Exploiting Perceptual Rules, Note Salience, and Melodic Smoothness," *Computer Music Journal*, Vol. 30, No. 4, pp. 80–98, 2006.

[4] S. Joo, S. Jo, and C. D. Yoo: "Melody extraction from polyphonic audio signal MIREX 2009," *MIREX Audio Melody Extraction Contest Abstracts*, 2009.

[5] R. Timmers and P. W. M Desain: "Vibrato: The questions and answers from musicians and science," In *Proc. Int. Conf. on Music Perception and Cognition*, 2000.