MELODY EXTRACTION FROM POLYPHONIC AUDIO SIGNAL MIREX2010

Sihyun Joo **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
s.joo@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) 2010 “Audio Melody Extraction” task. The proposed melody extraction algorithm can be divided into three steps: (1) a spectral analysis using a 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. Furthermore, a melody pitch range estimation is also considered to reduce false-positive and computation. In the third step, a single pitch sequence (melody line) is selected 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, octave mismatch, and dynamic variation in melody line [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 a 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 audio to find a suitable window length of each frame. Then, the short-time Fourier transform (STFT) with a 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. The effect of harmonic interference and octave mismatch can be reduced by considering several melody pitch candidates in each frame. Furthermore, a melody pitch range estimation is also considered to reduce false-positive and computation. In the third step, a single pitch sequence is identified from the melody pitch candidates based on general rules in a melody line.

2. METHOD DESCRIPTION

The melody extraction from a given polyphonic audio is performed in the STFT domain using the log frequency, otherwise known as *cent*. Frequency f_{Hz} in Hertz is converted to frequency f_{cent} in *cent* as follows:

$$f_{cent} = 6900 + 1200 \log_2 \frac{f_{Hz}}{440}. \tag{1}$$

This conversion formula divides one octave into 1200 *cent* and one note into 100 *cent*.

2.1 Spectral analysis with variable length window

The proposed algorithm uses a variable length window to reduce the effect of dynamic variation in melody line. Conventional melody extraction algorithms are performed the STFT with a fixed window length: Dressler used 46 ms window [5], Cancela used 92.9 ms window [6], and Ellis used 128 ms window [7]. Typically, aperiodic regions such as transient and *vibrato* regions should be analyzed by a short length window and vice versa for monotonous and periodic regions. Hence, finding a fixed window length appropriate for all types of audio may be impossible.

The window length is set based on the autocorrelation coefficient of the spectral magnitude of the polyphonic input audio: the autocorrelation coefficient is large during steady regions of a melody line and small during transient or *vibrato* regions. The autocorrelation coefficient ρ is defined as follows:

$$\rho_S(\tau, l) = \frac{\sum_k (|S(k,l)||S(k,l+\tau)|)}{\sqrt{\sum_k S^2(k,l)\ \sum_k S^2(k,l+\tau)}}, \tag{2}$$

where $S(k,l)$, l, τ, and k denote the Discrete Fourier Transform (DFT) of input audio, the current frame number, a distance from the current frame, and a bin number of DFT, respectively.

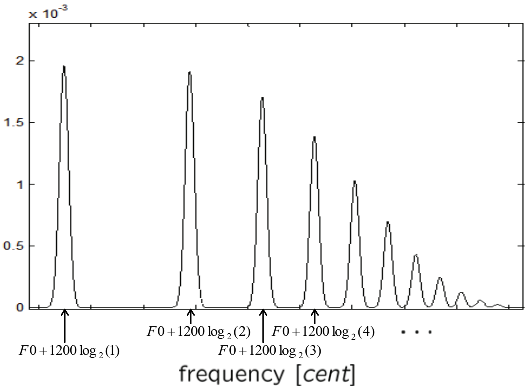


Figure 1. Harmonic structure model $H_\omega(k)$ ($H = 11$).

2.2 Pitch Candidate Estimation

In each frame, several melody pitch candidates are obtained to reduce the estimation errors due to harmonic interferences and octave mismatches. Here, the melody pitch candidate is defined as a possible estimate for the melody pitch. 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

$$H_\omega(k) = \sum_{m=1}^{H} A_m G(k; \omega + 1200 \log_2 m, W), \quad (3)$$

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

$$G(x; x_0, \varsigma) = \frac{1}{\sqrt{2\pi\varsigma^2}} \exp\left[-\frac{(x-x_0)^2}{2\varsigma^2}\right]. \quad (4)$$

Fig. 1 illustrates the harmonic structure model used in the considered algorithm.

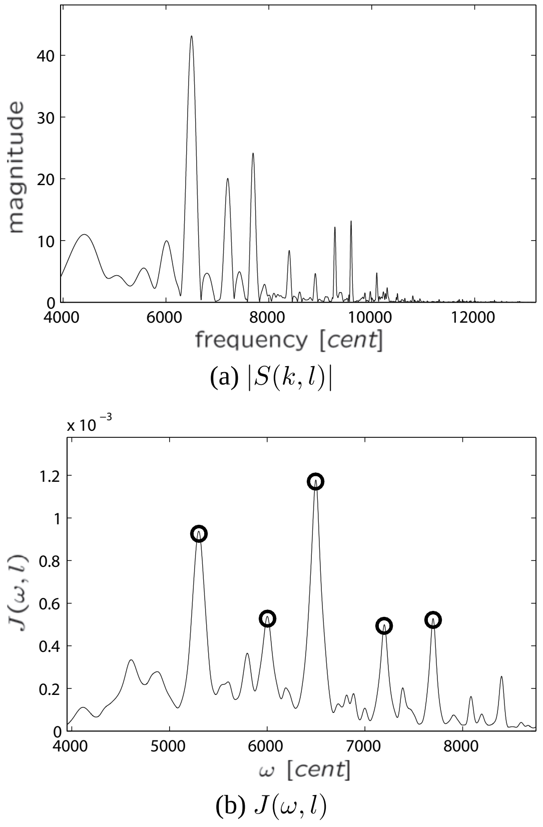
The weights are calculated as the inner-dot product between the harmonic structure with fundamental frequency ω and the spectral magnitudes. The weights of fundamental frequency ω in the l^{th} frame is mathematically expressed as

$$J(\omega, l) = \sum_k |S(k,l)| H_\omega(k). \quad (5)$$

Here, $J(\omega, l)$ informs the strength of the harmonic structure of a pitch frequency ω in the l^{th} frame. The pitch candidates are extracted as the peak values of $J(\omega, l)$. Fig. 2 (a) illustrates a certain STFT magnitude, and Fig. 2 (b) illustrates its $J(\omega, l)$. The circles in Fig. 2 (b) indicate the melody pitch candidates.

2.2.1 *Melody pitch range estimation*

It is well known that melody has the most dominant harmonic structure between 3950 *cent* and 8750 *cent*. However, many musicologist says that many music songs, especially in the popular music genre, consist of distinguishable parts called music structures [8], and note transitions are typically limited to an octave in each structure [1]. Thus, the search range for melody can be reduced if the polyphonic audio clip contains a single music structure. It gives effect to increase the accuracy of melody extraction and reduce the computation.



(a) $|S(k,l)|$

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

Figure 2. STFT magnitude and weights of l^{th} frame with fundamental frequency ω.

For estimating melody pitch range, reliable pitch candidates are collected: the pitch candidates which keep within 100 *cent* for 20 frames and more.Then, for estimating rough distribution for these candidates, a kernel density estimator with Gaussian kernel smoothing window [9] is used. It is wanted to know not the accurate distribution of the reliable pitch candidates but distribution tendency of melody pitch to set the search range for melody pitch. Thus, a large bandwidth is used for over-smoothing.

Fig. 3 shows the rough distribution of the reliable pitch candidates. Two bar lines in Fig.3 indicate two boundaries. If the other peak of which the distance from the largest peak position is shorter than an octave, then two boundaries are adjusted that they are allowed to include two peaks. These two boundaries are used as the new upper and lower boundaries of the search range for melody pitch extraction.

2.3 Pitch Sequence Identification

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

$$F_0(l) = \arg\max_\omega J(\omega, l). \quad (6)$$

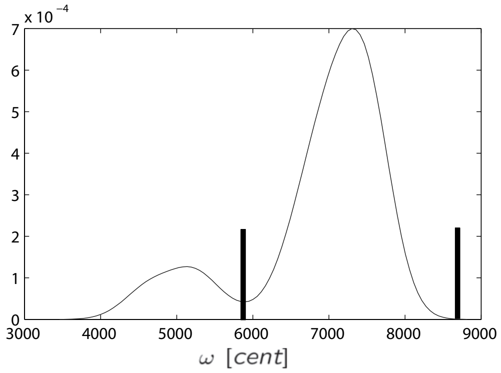


Figure 3. Rough distribution of reliable pitch candidates obtained by a kernel density estimator.

However, this result is not reliable, because of the three difficulties. Therefore, we consider several pitch candidates to get more reliable result.

N-best melody lines are estimated from N-best pitch candidates based on the following melody line properties:

1. The *vibrato* exhibits an extent of ± 60∼200 *cents* for singing voice and only ± 20∼30 *cents* for musical instruments [10].

2. The transitions are typically limited to an octave [1].

3. In general, a rest during singing is longer than 50 ms.

Several rules are defined based on the above properties to estimate N-best melody lines.

- Any two pitch candidates of successive frames are connected if the difference between the candidates is less than a threshold decided by the property 1.

- If any two pitch candidates of successive frames satisfied only property 2, transition is assumed to have occurred. The melody line is disconnected, and a new melody line starts. Frames between two disconnected points are defined as a *melody line segment*.

- If two inconsecutive frames with a time gap less than 50ms have pitch estimates that satisfy property 1, connect and interpolate between two pitch candidates.

- If pitch candidates do not satisfy property 1 and 2 for next few frames, then current melody is assumed absent.

A melody line candidate of each segment with the largest $J(\omega, l)$ is estimated as a melody line based on the melody definition: melody is a dominant pitch sequence of a polyphonic audio.

Once the pitch identification process is performed, any spurious pitch estimates are removed and replaced with a value interpolated between non-spurious estimates.

3. EVALUATION

3.1 Test database

The proposed algorithm was evaluated by using four database: Audio Description Contest (ADC) 2004, MIREX 2005, MIREX 2008, and MIREX 2009.

- ADC04: 20 audio clips of about 20 sec. from the following genres: daisy, jazz, opera, midi, and pop.

- MIREX05: 25 audio clips of 10∼40 sec. from the following genres: rock, rhythm and blues, pop, jazz, and solo classical piano.

- MIREX08: 8 audio clips of about 1 min. from following genre: north Indian classical vocal performances.

- MIREX09: 374 Karaoke recordings of 20∼40 sec. from Chinese pop. Three different levels of signal-to-accompaniment ratio (-5dB, 0dB, 5dB) are used for a total of 1122 audio clips.

All audio clips are single channel PCM with 16-bit quantization and 44.1 kHz sampling rate.

3.2 Evaluation Method

The reference frequencies of an unvoiced frame is considered as 0 Hz. The estimated pitch of a voiced frame will be considered correct when it satisfies the following condition:

$$|F_r(l) - F_e(l)| \le \frac{1}{4} tone \ (50 cent)$$

where $F_r(l)$ and $F_e(l)$ denote reference frequency and estimated pitch frequency of the l^{th} frame, respectively.

The performance of the proposed algorithm is evaluated with diverse aspects: Voicing Recall Rate (VR), Voicing False-Alarm Rate(VFA), Raw Pitch Accuracy (RPA), Raw Chroma Accuracy (RCA), and Overall Accuracy (OA) [1].

3.3 Result and Analysis

Table 1 shows overall results of melody extraction algorithms submitted to MIREX2010. JJY is the proposed algorithm in this paper. The result of each database is not summarized because the size, genre, and musical feature of each database are different. In addition, the experiment condition of each database is also different: the results of MIREX09 database are obtained in diverse signal-to-accompaniment ratios, but the results of the other database are not. In table 1, the best result of each performance measure is written in bold.

The proposed algorithm shows a quite good performance on the four different database. The algorithm achieves the best OA, RPA, and RCA on the ADC04 and the MIREX08 database. The RCA of the MIREX05 database is also the highest. The OA and the RPA of the MIREX05 database rank second, but differences against the best results are just 0.6 percent and 0.1 percent, respectively. Likewise, the RPA and the RCA of the MIREX09(0dB) database has

Algorithm	OA	RPA	RCA	VR	VFA
JJY	**71.9**	**79.6**	**85.3**	**93.7**	50.1
HJ	61.3	76.8	79.8	73.4	**21.0**
SG	69.9	75.2	78.2	80.6	23.2
TOOS	53.7	62.9	73.4	79.6	32.8

(a) ADC04 evaluation result.

Algorithm	OA	RPA	RCA	VR	VFA
JJY	61.5	71.6	**78.6**	**97.3**	70.2
HJ	53.9	**71.7**	74.9	70.8	44.9
SG	**62.1**	61.8	73.7	76.4	**22.8**
TOOS	60.8	68.9	74.6	84.7	41.9

(b) MIREX05 evaluation result.

Algorithm	OA	RPA	RCA	VR	VFA
JJY	**79.6**	**88.6**	**90.4**	**95.1**	44.6
HJ	76.8	86.0	86.8	89.8	**22.8**
SG	77.7	85.7	86.9	86.2	23.2
TOOS	72.0	82.4	86.2	85.5	23.1

(c) MIREX08 evaluation result.

Algorithm	OA	RPA	RCA	VR	VFA
JJY	62.9	82.2	84.6	**98.3**	70.7
HJ	**76.2**	**83.2**	84.2	82.1	**14.3**
SG	73.6	80.1	85.5	89.7	30.2
TOOS	72.2	82.6	**86.2**	94.2	38.6

(d) MIREX09(0dB) evaluation result.

Table 1. Evaluation result of ADC04, MIREX05, MIREX08, MIREX09(0dB) database (Unit:%).

less than 1.6 percent difference against the highest accuracy. These results indicate that the proposed algorithm is not biased toward certain type or genre of audio data.

In spite of a good performance of the RPA, the OA result of the MIREX09(0dB) database is quite low due to the high VFA. Table 1 (a) and (b) show that the difference between VR and VFA must be larger to get a higher OA. In table 1 (a), SG has lower RPA than the RPA of HJ, but the OA is higher. In addition, SG shows the best OA despite of the lowest RPA due to the largest difference between the VR and VFA in table 1 (b). Hence, an algorithm increasing the difference has to be created to improve the OA.

4. CONCLUSION AND FUTURE WORKS

The proposed algorithm for the MIREX 2010 audio melody extraction task is described in this paper. The algorithm consists of three steps. First, a spectral analysis is performed by using a variable length window. Second, melody pitch candidates of each frame are obtained and a melody pitch range is estimated. Third, a single pitch sequence is estimated from possible pitch sequences.

As part of our continuing research, we plan to focus on these problems because the melody extraction techniques can be used for a music information retrieval (MIR) system or a plagiarism audio search system.

5. 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] K. Dressler: “An Auditory Streaming Approach on Melody Extraction,” In *MIREX Audio Melody Extraction Contest Abstracts*, 2006.

[6] P. Cancela: “Tracking melody in polyphonic audio. MIREX 2008,” In *MIREX Audio Melody Extraction Contest Abstracts*, 2008.

[7] D. P. W. Ellis and G. E. Poliner: “Classification-based melody transcription,” In *Machine Learning*, Vol. 65, pp. 439–456, 2006.

[8] J. Paulus and A. Klapuri: “Music Structure Analysis by Finding Repeated Parts,” In *Proceedings of the 1st Audio and Music Computing for Multimedia Workshop (AMCMM2006)*, pp. 59–68, 2006.

[9] A. W. Bowman and A. Azzalini: *Applied Smoothing Techniques for Data Analysis*, Oxford University Press, 1997.

[10] 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.