Melody pitch estimation based on range estimation and candidate extraction using harmonic structure model

Seokhwan Jo, Sihyun Joo, Chang D. Yoo

Department of Electrical Engineering, KAIST,
373-1 Guseong-dong, Yuseong-gu, Daejeon, 305-701, Korea
antiland00@kaist.ac.kr, s.joo@kaist.ac.kr, cdyoo@ee.kaist.ac.kr

Abstract

This paper proposes an algorithm to estimate the melody pitch line (the most dominant pitch sequence) of a given polyphonic audio based on melody range estimation and pitch candidate extraction using a harmonic structure model similar to that proposed by Goto. This paper defines melody pitch candidate as a list of pitch candidates that produces the best-fit harmonic models to the polyphonic audio. In many melody extraction algorithms proposed in the past, multiple-pitch extractor (MPE) is often performed for extracting melody pitch candidates; however, the MPE serves the purpose of estimating all pitches within a frame of a polyphonic audio and does not necessarily provide melody pitch candidates. The estimated weights of the harmonic structure model which must be obtained for extracting the pitch candidates are liable to octave error and strong low frequency interference, and therefore, certain refinement after the estimation must be performed. As a refinement, the algorithm measures the degree of harmonic fitness of each candidate. Furthermore, a melody pitch range is estimated to reduce false-positive pitch candidates. The melody pitch range is estimated based on the distribution of the best pitch candidates with long duration. Experimental results show that the proposed extraction algorithm performed better than many of the algorithms proposed in the past.

Index Terms: melody extraction, pitch candidates, harmonic structure model, pitch candidate refinement, melody pitch range

1. Introduction

Although the debate on the definition of melody is ongoing [1, 2, 3], many experts concur that melody should be the dominant pitch sequence of a polyphonic audio, and experts believe that people recognize music by its melody. For this reason, melody extraction from polyphonic audio is playing an important role in music information retrieval (MIR), and melody extraction has the potential to be used in contents-based music retrieval, determining audio plagiarism, and music transcription.

Many melody extraction algorithms perform multiple-pitch extraction (MPE) in determining all the possible pitch candidates of the melody. Various multiple-pitch extraction techniques have been proposed in recent years, albeit with limited success. Goto estimated the weights of prior tone-models over all possible fundamental frequencies (F0s) by peak-picking the short-time Fourier transform (STFT) [2]. The multiple-pitch were extracted from these weights. Paiva found the possible pitch from the largest peaks of the STFT [3]. Klapuri estimated multiple-pitch by calculating the weighted sum of the magnitudes of its harmonic partials [4].

In this paper, a simple but effective melody pitch estimation algorithm based on melody range estimation and pitch candidate extraction using harmonic structure model is proposed. This paper defines melody pitch candidate as a list of pitch candidates that produces the best-fit harmonic models to the polyphonic audio. Unlike the purpose of the MPE, which is to extract all pitches in the polyphonic audio, the purpose of melody pitch candidates extraction is to extract possible pitch candidates of the melody. To extract pitch candidates, the weights of the harmonic structure model similar to Goto's model [2] are estimated in the STFT magnitude domain. Since the extracted pitch candidates obtained from these weights are liable to octave error and strong low frequency interference, certain refinement after the estimation must be performed. As a refinement, the algorithm measures the degree of harmonic fitness of each candidate.

In this paper, the range of melody pitch is estimated to enhance the accuracy of melody pitch extraction. It is well known that melody has the most dominant harmonic structure in the middle and high frequency regions [2], and thus, most melody extraction algorithms set search range for the melody between 80Hz and 1280Hz in frequency domain [1, 2, 3, 5]. However, this range is too large and it often leads to octave error in pitch estimation. For this reason, the melody pitch range is estimated based on the distribution of the best pitch candidates with long duration.

This paper is organized as follows. Section 2 presents the melody pitch candidates extraction algorithm. Section 2.1 presents an algorithm to extract pitch candidates from the weights of the harmonic structure model. Section 2.2 presents an algorithm to refine the estimates of the pitch candidates. Section 2.3 presents an algorithm for estimating the melody pitch range. Section 3 provides experimental results. Finally, Section 4 concludes this paper.

2. Melody pitch candidates extraction

This paper proposes an algorithm for extracting melody pitch from a given polyphonic audio. The extraction 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:

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

This conversion formula divides an octave into 1200 *cent* and a note into 100 *cent*. The proposed algorithm for pitch candidate extraction is presented in the following sub-sections.

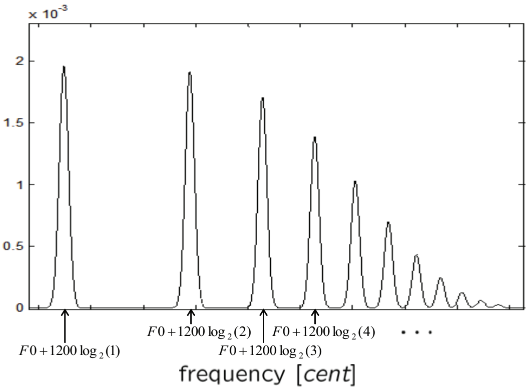


Figure 1: Harmonic structure model $H_\omega(k)$ of which $H = 11$.

2.1. Melody pitch candidates extraction using harmonic structure model

To extract pitch candidates from polyphonic audio, the weights of the harmonic structure model similar to that proposed in [2] are estimated. In this paper, the harmonic structure model is mathematically defined as

$$H_\omega(k) = \sum_{m=1}^{H} A_m G(k; \omega + 1200\log_2 m, W), \qquad (2)$$

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

$$G(x; x_0, \varsigma) = \frac{1}{\sqrt{2\pi\varsigma^2}} \exp\left[-\frac{(x - x_0)^2}{2\varsigma^2}\right]. \qquad (3)$$

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

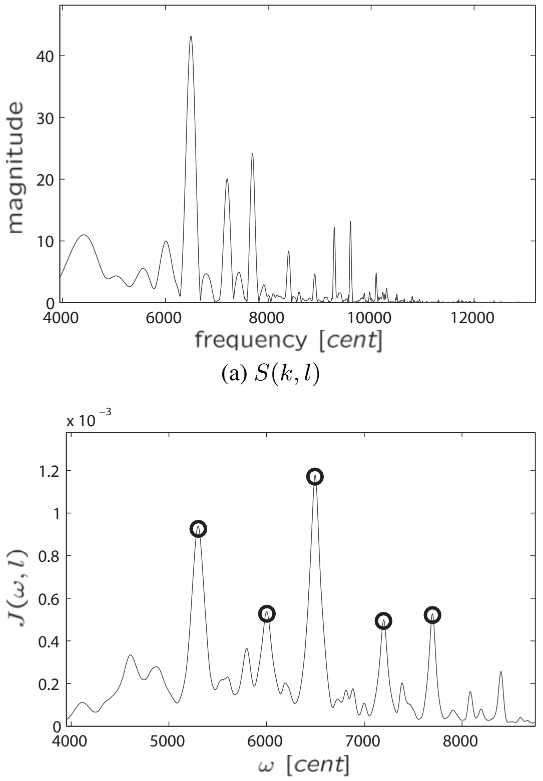
The weights are calculated as the inner-dot product between the harmonic structure with fundamental frequency ω and the spectral magnitudes of the polyphonic audio. The weight of fundamental frequency ω in the l^{th} frame is mathematically expressed as

$$J(\omega, l) = \sum_k |S(k, l)| H_\omega(k), \qquad (4)$$

where $S(k, l)$ is the k^{th} coefficient of the STFT of the l^{th} frame. Here, $J(\omega, l)$ indicates the strength of the harmonic structure with pitch frequency ω in the l^{th} frame. The pitch candidates are the frequency where $J(\omega, l)$ peaks. 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. Pitch candidates refinement

The pitch candidates can be extracted from $J(\omega, l)$; however, these pitch candidates are liable to octave error and strong low frequency interference. Thus, a refinement procedure is required: the algorithm measures a degree of harmonic fitness of each candidate and retains those candidates with the highest harmonic fitness. Specifically, the extracted pitch candidates using $J(\omega, l)$ are refined by considering the number of harmonics that includes a spectral peak within a possible harmonic peak range. The modified weight taking into consideration the harmonic fitness is mathematically expressed by



(a) $S(k, l)$

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

Figure 2: $S(k, l)$ and its $J(\omega, l)$. The circles (o) indicate the melody pitch candidates.

$$J'(\omega, l) = \frac{h_\omega}{H} J(\omega, l), \qquad (5)$$

where h_ω is the number of harmonic partials detected correctly. The m^{th} harmonic partial is considered correctly detected when the spectral peak of large magnitude exists in the range $\eta_{\omega,m}$. Here, $\eta_{\omega,m}$ is a range of frequency bins in the vicinity of the m^{th} harmonic partial when $F0 = \omega$, and it can be expressed mathematically as

$$\eta_{\omega,m} = [\omega + 1200\log_2 m - 50, \quad \omega + 1200\log_2 m + 50]. \ (6)$$

The weight of the harmonic structure model of each pitch candidate is updated by Eq. (5) based on the harmonic fitness. Then, the refined pitch candidates are obtained from $J'(\omega, l)$.

2.3. Melody pitch range estimation

Melody is considered 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 [6], 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. If the search range is estimated adequately for a given audio clip, then it is possible to increase the accuracy of melody extraction and reduce the computation in the search. Melody

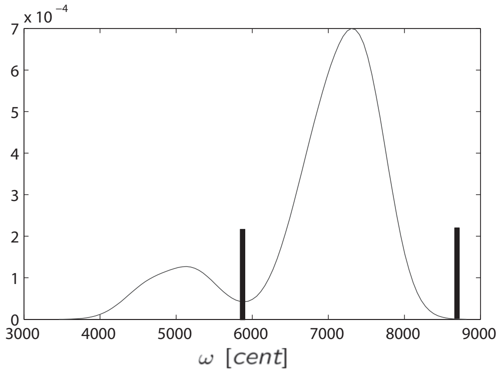


Figure 3: Rough distribution of reliable pitch candidates obtained by a kernel density estimator.

line of the database for evaluation in this paper does not exceed two octaves.

To reliably estimate the melody pitch range, the distribution of reliable pitch candidates is considered. A reliable pitch candidate is defined as a steady pitch candidate whose pitch value remains fairly constant (± 100 *cent*) for more than 20 frames. In estimating the distribution, a kernel density estimator based Gaussian kernel smoothing window [7] is used. Fig. 3 shows an example of the distribution of the reliable pitch candidates. With high probability, the pitch candidates are assumed to lie between two valley positions of the distribution. The valley positions are taken such that probability that the candidates lie between the two position are very high.

The mean and standard deviation of the pitch candidates that lie between the valley position are estimated, and based on these statistics, the lower and upper bounds of the search range are set. The lower and upper bounds are given as follows:

$$F_l = \max(3950, \mu - 3\sigma), \tag{7}$$

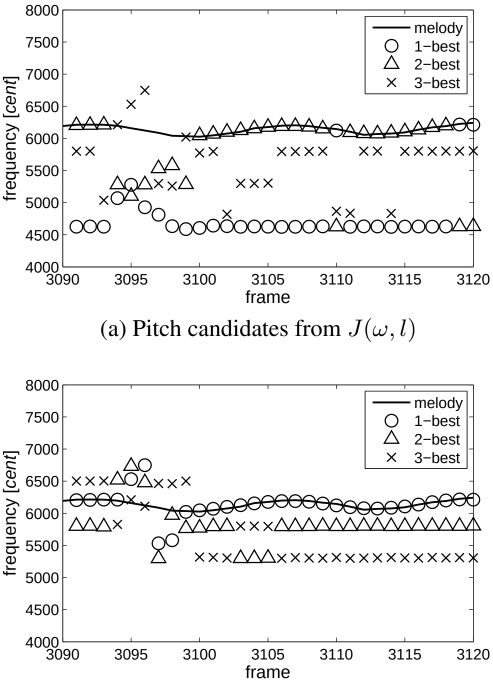
$$F_h = \min(8750, \mu + 3\sigma), \tag{8}$$

where μ and σ are mean and standard deviation of the reliable pitch candidates between the valley positions, respectively.

Pitch candidates and true melody line are shown as Fig. 4. It is shown that the refinement and melody range estimation improve the accuracy of pitch candidates (Fig. 4 (b)).

3. Evaluation

The proposed algorithm was evaluated using two database. The first database is obtained from the audio melody extraction task of Audio Description Contest (ADC) 2004, and it contains 20 polyphonic musical audio clips. The second one is obtained from the audio melody extraction task of Music Information Retrieval Evaluation eXchange (MIREX) 2005, and it contains 13 clips. All audio clips are single channel PCM data with 44.1kHz sampling rate and 16-bit quantization. The ADC04 database consists of more nonvocal melody than vocal melody. The MIREX05 database consists of more vocal melody. The search range of ω in Eq. (4) was set between 3950 *cent* and 8750 *cent*. Hanning window was used with variable frame lengths (32ms, 64ms, 128ms) and 10ms frame hop size. The harmonic structure parameters were set as follows: $H = 11$, $W = 15$, $A_m = G(m; 1, 0.3H)$. The proportion of frames with true melody pitch outside the bounds set by Eq. (7) and Eq. (8) is 0.0342%.



(a) Pitch candidates from $J(\omega, l)$

(b) Pitch candidates after refinement and range estimation

Figure 4: Melody pitch candidates.

The performance of the proposed algorithm was evaluated in terms of raw pitch accuracy (RPA) and raw chroma accuracy (RCA) [1]. The estimated pitch is considered correct when the absolute value of the difference between the reference frequency and the estimated pitch frequency is less than 50 *cent*. The RPA is defined as the proportional frames with the correct pitch estimates. The RCA is defined in the same manner as the RPA; however, both the estimated and reference frequencies are mapped into a single octave in order to forgive octave transpositions.

Table 1 gives the RPA score of pitch candidates (RPA_PC) for different frame length and database. To compare the performance of the melody extraction using melody pitch candidates, the following method was used. Melody pitch was estimated roughly to find the ω that maximizes $J'(\omega, l)$:

$$F0(l) = \arg\max_{\omega} J'(\omega, l), \quad F_l \leq \omega \leq F_h. \tag{9}$$

Afterwards, any spurious melody pitch estimates were removed and replaced with other pitch candidates close to non-spurious estimates.

The melody extraction using these candidates was compared to the other famous melody extraction algorithms such as algorithms proposed by Goto [2], Paiva et al. [3], Ryynänen el al. [8], and Ellis et al. [9]. Their performances are based on results of the Music Information Retrieval Evaluation eXchange (MIREX) [10]. The ADC 2004 database was used for the performance comparison of melody extraction. Table 2 shows the evaluation results for all algorithms considered. The melody extraction using the melody pitch candidates obtained from the proposed algorithm outperformed the others in terms of the RPA and the RCA.

Table 1: *RPA_PC.*

	frame length	ADC 2004	MIREX 2005
1-best	32ms	79.2%	79.7%
	64ms	79.9%	79.4%
	128ms	77.3%	76.4%
3-best	32ms	90.3%	88.7%
	64ms	91.3%	90.9%
	128ms	90.1%	89.8%
5-best	32ms	92.3%	90.4%
	64ms	93.9%	92.9%
	128ms	92.8%	92.4%

Table 2: *Melody extraction results. The PA means the proposed algorithm. The number in parentheses is the year when their algorithms were submitted to the MIREX.*

	RPA	RCA
Goto [2]	65.8% (2005)	71.8% (2005)
Paiva el al. [3]	62.7% (2005)	66.7% (2005)
Ryynänen el al. [8]	68.6% (2005)	74.1% (2005)
Ellis el al. [9]	73.2% (2006)	76.4% (2006)
PA (frame length: 64ms)	80.8%	86.7%

4. Conclusion

In this paper, an algorithm to estimate the melody pitch line from a given polyphonic audio based on melody range estimation and pitch candidate extraction using a harmonic structure model is proposed. Melody pitch candidate is defined as a list of pitch candidates that produces the best-fit harmonic models to the polyphonic audio. Most melody extraction algorithms executes the MPE for extracting melody pitch candidates; however, the MPE serves the purpose of estimating all pitches within a frame of a polyphonic audio and does not necessarily provide melody pitch candidates. The estimated weights of the harmonic structure model which must be obtained for extracting the pitch candidates are liable to octave error and strong low frequency interference, and therefore, certain refinement after the estimation must be performed. As a refinement, the algorithm measures the degree of harmonic fitness of each candidate. Furthermore, a melody pitch range is estimated to reduce false-positive pitch candidates. The melody pitch range is estimated based on the distribution of the best pitch candidates with long duration. Experimental results show that the proposed extraction algorithm outperformed the other famous melody extraction algorithms.

5. Acknowledgements

This work was supported by the Ministry of Culture, Sports and Tourism (MCST) and Korea Culture Content Agency (KOCCA) in the Culture Technology (CT) Research and Development Program 2009.

6. References

[1] Poliner, G. E., Ellis, D. P. W., and Ehmann, A. F., "Melody transcription from music audio: approach and evaluation", *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1247–1256, 2007.

[2] Goto, M., "A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals", *Speech Communication*, 43(4):311–329, 2004.

[3] Paiva, R. P., Mendes, T., and Cardoso, A., "A methodology for detection of melody in polyphonic music signals", *AES 116th Convention*, 2004.

[4] Klapuri, A., "Multiple fundamental frequency estimation by summing harmonic amplitudes", *Proceedings of the International Conference on Music Information Retrieval 2006*, 216–221, 2006.

[5] Dressler, K., "Audio melody extraction for MIREX 2009", *MIREX 2009 Audio Melody Extraction Contest*, 2009.

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

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

[8] Ryynänen, M. P. and Klapuri, A. P., "Note event model- ing for audio melody extraction", *MIREX 2005 Audio Melody Extraction Contest*, 2005.

[9] Ellis, D. P. W. and Poliner, G. E., "Classification-based melody transcription", *Machine Learning*, 65:439–456, 2006.

[10] Downie, J. S., West, K., Ehmann, A., and Vincent E, "The 2005 music information retrieval evaluation exchange (mirex 2005): preliminary overview", *Proceedings of the Sixth International Conference on Music Information Retrieval*, 320–323, 2005.