MELODY EXTRACTION FROM POLYPHONIC AUDIO SIGNAL MIREX2009

Sihyun Joo **Seokhwan Jo** **Chang D. Yoo**
Div. of EE, Dept. of EECS
Korea Advanced Institute of Science and Technology
373-1 Guseong Dong, Yuseong 1Gu, Daejeon 305-701, Korea
s.joo@kaist.ac.kr antiland@kaist.ac.kr cdyoo@ee.kaist.ac.kr

ABSTRACT

This paper describes the proposed algorithm submitted to the MIREX 2009 "Audio Melody Extraction" task. The algorithm addresses the task of extracting the predominant melody pitch from a polyphonic audio signal. The algorithm extracts the melody pitch in three steps. In the first step, transient analysis is performed on the polyphonic audio signal to determine the analysis frame length, and then a fixed number of pitch candidates are obtained by ranking the weights of the harmonic structure of the windowed signal. In the second step, a single dominant pitch sequence (melody line) is selected from the many possible pitch sequences based on the following properties of melody line: (1) while the nominal dynamic range of a singing *vibrato* is $\pm$ 60$\sim$200 *cents*, it is only $\pm$ 20$\sim$30 *cents* for instruments; (2) melody transitions are typically limited to one octave; (3) a rest during singing is often longer than 50ms. In the third step, a smoothing process is performed to refine the estimated pitch sequence.

Keywords: MIREX, Variable length window, Harmonic structure model, Predominant pitch

1. INTRODUCTION

We have seen tremendous progress in the area of melody extraction over last decade. Certainly, the MIREX audio melody extraction contest has had considerable impact on the progress.

Nevertheless, there is still no clear definition of melody. Melody is defined in many different ways. Solomon (1997) defines melody as, *a combination of a pitch series and a rhythm having a clearly defined shape* [3]. Goto (2004) defines melody as, *the most predominant pitch supported by harmonics within an intentionally limited frequency range* [4]. Paiva (2007) defines melody as, *the dominant individual pitched line in a musical ensemble* [6]. Even though there are diverse definitions of melody, most definitions commonly refer melody as the dominant pitch sequence of a polyphonic audio.

With this definition of melody in mind, melody extraction is difficult for the following reasons:

- Harmonic interference: The Harmonics of subdominant melodies tend to act as noise and can interfere in the estimation of the harmonics of the predominant melody.
- Octave mismatch: The estimated pitch can be one octave higher or lower than the ground-truth.
- Dynamic variation in melody: Accurate pitch estimation in the beginning, end and sudden transient regions of a melody can be difficult.

The above difficulties can be partially alleviated by considering several pitch candidates in each frame and using a variable length window. Considering several pitch candidates (N-best pitch candidates) in each frame instead of just one pitch to estimate the melody pitch (ground-truth) reduces the possibility of incorrect estimation especially when the melody pitch at a particular frame is not the most dominant pitch. The use of a variable length window allows automatic time-frequency adjustment in analyzing the pitch: low pitch frequency should be analyzed by a long window and vice versa for high pitch frequency.

The overall block diagram of proposed algorithm is shown in Fig.1. The proposed algorithm extracts the pitch sequence (melody line) in three steps: (1) pitch candidate estimation, (2) pitch sequence identification, and (3) smoothing process. First, a transient analysis is performed on the polyphonic input audio to find a suitable frame length. The N-best pitch candidates are determined by ranking the weight of the harmonic structure of each windowed frame. Second, the melody line is estimated from the N-best pitch candidates of each frame based on a general rule on the melody line. Third, a smoothing process is performed to refine any spurious pitch estimates.

2. METHOD DESCRIPTION

2.1 Pitch Candidate Estimation

Conventional melody extraction algorithms are based on fixed window length: Paiva used 20 ms window [7], Dressler

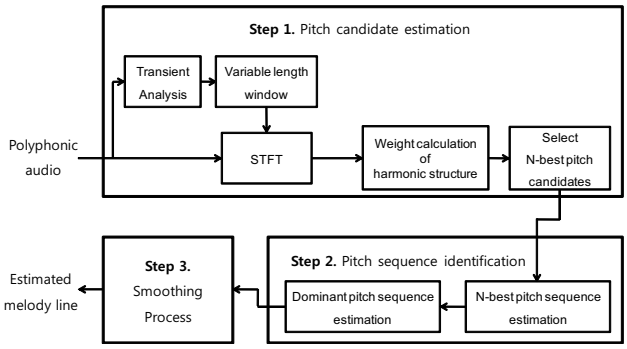

Figure 1. Overall system block diagram.

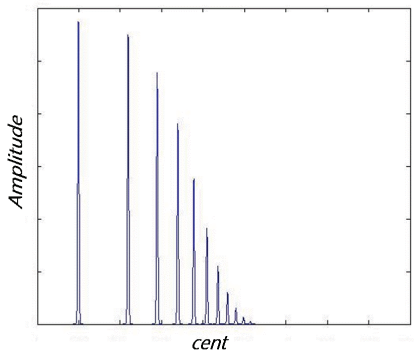

Figure 2. Harmonic structure model.

used 46 ms window [2], Cancela used 92.9 ms window [5], and Poliner used 128 ms window [1]. Typically, a long window which provides high frequency resolution but low temporal resolution would be more appropriate for monotonous and periodic region of a melody. In contrast, a short length window would be more appropriate for aperiodic regions such as transient and vibrato regions. Therefore, finding one fixed window length appropriate for all types of audio may be impossible.

The proposed algorithm uses a variable length window to capture the time-varying characteristic of a melody line. The window length is set based on the autocorrelation coefficient of the Fourier transform magnitude of the polyphonic input audio: the autocorrelation coefficient is large during the steady regions of the melody line and small during transient or *vibrato* regions. The autocorrelation coefficient of signal X at the l^{th} frame is approximated as follows:

$$\rho_X^l(\tau) = \frac{\sum_k |X_l(k)||X_{l+\tau}(k)|}{\sqrt{\sum_k X_l^2(k) \ \sum_k X_{l+\tau}^2(k)}} \quad (1)$$

where $X_l(k)$ denotes the k^{th} coefficient of the discrete Fourier transform (DFT) of the l^{th} frame.

When the window length is too short, the frame will not contain enough periods to accurately estimate the melody pitch. On the other hand, a long window would be inappropriate for transition. For this reasons, a minimum and maximum window lengths are set.

In the proposed algorithm, a fixed number of pitch candidates are obtained to reduce the estimation errors due to harmonic interferences and octave mismatches. The N-best pitch candidates are estimated according to the rank of the weight of the harmonic structure. The weight of each pitch candidate is determined by the weight of the harmonic structure pertaining to the pitch frequency in the framed signal. Fig.2 illustrates the harmonic structure model used, which is a modified version of the model used in [4].

Due to radiation effects, the weight on each pitch candidate is biased towards low pitch frequency. To compensate for this, the weight of a harmonic structure model is modified by multiplying the following prior function:

$$f(x) = 1 - e^{\frac{(x-2750)^2}{2c}} \quad (2)$$

where x, c denote the pitch candidates - it is assumed that melody line exists between $2750cent$ and $7550cent$ ($80Hz$ and $1280Hz$, 4 octaves) range - and a constant which controls the shape of the function, respectively.

2.2 Pitch Sequence Identification

Once the N-best pitch candidates are obtained for all frames, a single pitch sequence that best represents the melody is identified. This identification process can be performed in two ways: the probability-based way and the rule-based way [1]. The probability-based way generally incorporates a probabilistic model such as the hidden Markov model (HMM). The rule-based way is implemented by a rule defined based on the properties of a melody. Compared to the probability-based way that have been proposed sofar, the rule-based way generally do better to minimize the difference between Raw Pitch Accuracy and Chroma Accuracy [1]. The average accuracy is also better. The proposed pitch identification algorithm that estimates the pitch sequence (melody line) can be considered as a rule-based method.

The proposed algorithm estimates melody line from N-best pitch candidates based on the following three properties of the melody:

1. While the nominal dynamic range of a singing *vibrato* is $\pm$ 60$\sim$200 *cents*, it is only $\pm$ 20$\sim$30 *cents* for instruments [8].

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

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

2.3 Smoothing Process

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.

Table 1. MIREX 2009 Audio Melody Extraction dataset.

Dataset	Melody	Number of files
ADC04	Vocal melody	8
	Nonvocal melody	12
MIREX05	Vocal melody	16
	Nonvocal melody	9
MIREX08	Vocal melody	8
MIREX09	Vocal melody	374

Table 2. MIREX 2009 Audio Melody Extraction overall summary results of our algorithm - Equal dataset weight.

Vx Recall	Vx False Alm	RPA	RCA	OA
61.0%	29.4%	73.3%	79.7%	56.6%

3. EVALUATION

3.1 IMPLEMENTATION

The proposed melody extraction algorithm has been implemented in Matlab. The implemented algorithm runs on Window Matlab version 2007 or more updated versions.

3.2 Test Dataset

Four CD-quality (16-bit quantization, 44.1 kHz sample rate) test datasets are used for the MIREX 2009 Audio Melody Extraction task evaluation. Table 3.2 shows the organization of the overall dataset.

ADC04 and MIREX05 test datasets consist of vocal melody and nonvocal melody data. On the other hand, MIREX08 and 09 datasets consist of only vocal melody data. MIREX09 dataset is mixed with three different (+5dB, 0dB, -5dB RMS) melodic voice.

3.3 Evaluation Rule and Method

The estimated pitch of a voiced frame will be considered correct when the absolute value of the difference between the reference frequency and the estimated pitch frequency is less than quarter tone (50 cent). This is mathematically shown by the following:

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

where $F_r(i)$ and $F_e(i)$ denote reference frequency and estimated pitch frequency of the i^{th} frame, respectively. The reference frequency of an unvoiced frame considered as 0 Hz.

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

Table 3. MIREX 2009 Audio Melody Extraction overall average results of RPA. Unweighted average is the average of datasets with equal weight. Weighted average is the average of datasets weighted by the number of data.

Participant	Unweighted RPA Avg.	Weighted RPA Avg.	Overall RPA Avg.
Dressler	80.6%	77.6%	79.1%
Tachibana	75.1%	80.1%	77.6%
Joo	**73.3%**	**73.0%**	**73.2%**
Durrieu1	74.5%	68.6%	71.5%
Wendelboe	73.4%	66.4%	69.9%
Rao	72.2%	67.4%	69.8%
Durrieu2	72.1%	65.3%	68.7%
Hsu1	66.1%	68.4%	67.3%
Cao1	63.5%	58.9%	61.2%
Cao2	63.5%	58.9%	61.2%
Cancela	64.1%	51.8%	57.9%
Hsu2	51.1%	50.4%	50.8%

Table 4. MIREX 2009 Audio Melody Extraction overall average results of RCA.

Participant	Unweighted RCA Avg.	Weighted RCA Avg.	Overall RCA Avg.
Tachibana	80.3%	83.8%	82.1%
Dressler	82.5%	79.5%	81.0%
Joo	**79.7%**	**77.7%**	**78.7%**
Durrieu1	76.8%	71.4%	74.1%
Wendelboe	77.5%	70.3%	73.9%
Rao	76.3%	70.5%	73.4%
Durrieu2	75.7%	69.8%	72.8%
Hsu1	72.6%	72.1%	72.3%
Hsu2	67.1%	67.0%	67.0%
Cao1	66.3%	62.8%	64.5%
Cao2	66.3%	62.8%	64.5%
Cancela	65.8%	54.2%	60.0%

3.4 Result

The overall results of our melody extraction algorithm submitted to MIREX 2009 are shown in Table 3.4. The RPA and RCA results are obtained by setting the weight of the database both equal and also according to the number of files in the database. It should be noted that the number of files in MIREX09 is over 10 times that of the sum of all the other database. The musical feature and genre of database of each year are different. Hence, equal weight has to be allocated to figure out the general performance of the algorithm about diverse kind of database.

We note that our proposed algorithm ranks 3/12 for both overall average of raw pitch and chroma accuracy (See Table 3.4 and Table 3.4). The raw pitch and chroma accuracy, both of them have very small difference between the unweighted and weighted average compared to other results. It means that the pitch estimation algorithm is not biased

toward certain type of music or data. If an algorithm is biased, the variation of each dataset results will be very large. It affects the large difference between the unweighted and weighted average.

The overall accuracy of our algorithm ranked 6th due to the poor performance of the voice detection algorithm, which is based on simple adaptive energy threshold. Voicing detection errors lead to high false negative (voiced frames detected as unvoiced frames) and false positive (unvoiced frames detected as voiced frames). This influences the overall accuracy considerably. As Tachibana, Wendelboe and many others [4], we focused more on raw pitch and chroma accuracy, and not the overall accuracy. In conclusion, the overall accuracy of our system can be increased by improving the voice detection algorithm.

4. CONCLUSION AND FUTURE WORKS

The proposed algorithm for the MIREX 2009 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, and the N-best pitch candidates are selected based on the weight of a harmonic structure model at each frame. Second, a single pitch sequence is estimated from the pitch candidates by using rule-based method. Third, a smoothing process replaces any spurious pitch estimates with a value interpolated between non-spurious estimates.

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. ACKNOWLEDGEMENTS

The authors would like to thank to the IMIRSEL team at the University of Illinois at Urbana-Champaign (UIUC) for arranging and running the MIREX 2009 evaluations.

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

[3] L. Solomon: *Music theory glossary*,Web publication, last updated 2002.

[4] 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.

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

[6] 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.

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

[8] 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.