# Extracting Predominant Melody of Polyphonic Music based on Harmonic Structure

**Jea-Yul Yoon[1], Chai-Jong Song[1,2], Seok-Pil Lee[2] and Hochong Park[1]**

[1]Dept. of Electronics Engineering,
Kwangwoon University
Seoul, Korea
{yoon, hcpark}@kw.ac.kr

[2]Digital Media R&D Center
Korea Electronics Technology Institute
Seoul, Korea
{jcsong, lspbio}@keti.re.kr

## ABSTRACT

In this paper, we propose a method for extracting predominant melody of polyphonic music based on harmonic structure. We extract all meaningful spectral peaks contained in the polyphonic signal and select the F0 candidates by verifying the required condition of harmonic structure. Then, we determine the predominant F0 by running the pitch tracking based on the rank of F0 candidates which is computed from the average energy of harmonic peaks. In addition, we use a post-processing module which corrects the pitch doubling and halving errors by analyzing the group characteristics of F0s.

## 1. INTRODUCTION

In order to analyze and search musical data efficiently, there has been growing interest in CASA (computational auditory scene analysis), QbH (query by humming) and MIR (music information retrieval). These tasks commonly require the extraction of predominant melody or vocal melody from polyphonic music, and a great deal of research is being  carried out to this end. The current technologies can be classified into the decomposition method for spectral parameters and the modeling method using the statistical properties [1].

The spectral decomposition method extracts the melody without probability modeling of signal by using spectral harmonics, frequency, timbre characteristics, NMF (non-negative matrix factorization) and filter bank. The existing technologies so far have differences in that they implement the same goal in different ways. The method proposed in this paper can search the fundamental frequency (F0) more accurately since it obtains all F0 candidates using all given spectral peak positions. In addition, the proposed method is a simple and unsupervised method, and shows high accuracy by using the harmonic structure of spectrum without complex processing of signal modeling.

## 2. PROPOSED MELODY EXTRACTION METHOD

### 2.1 F0 Candidates Extraction

Since polyphonic music contains multiple sound sources at the same time, the extraction of main melody needs to search for the multi-pitch frequencies and to select a F0 corresponding to the main melody out of the extracted multi-pitch frequencies. In order to implement these processes effectively, in this paper, we propose a multi-pitch extraction and main melody extraction method based on the harmonic structure that is the important characteristics of musical signals.

Since the fundamental frequency of music signal is determined by the low-frequency component, the input music signal is down-sampled to 8 kHz sampling frequency in the pre-processing module. Multi-pitch extraction module consists of the extraction of pitch information included in the signal, the selection of valid pitch according to the availability and the accuracy of harmonic structure of the extracted pitch, and the computation of rank of each pitch.

This method first determines all meaningful spectral peaks contained in the signal. Then, by selecting the spectral peaks with a high probability of harmonic peak and analyzing the gap between the peak positions, it checks if the selected peaks meet the conditions of harmonic structure and chooses the F0 candidates satisfying these conditions. It combines the harmonic components of each F0 candidate and generates the harmonic groups. Finally, it determines the priority rank of each F0 candidate by calculating the average energy of each harmonic group.

### 2.2 Pitch Tracking

We select a predominant F0 through the pitch tracking of F0 candidates extracted in each frame. The pitch tracking module runs the pitch tracking by considering the F0 continuity between the adjacent frames and by considering the rank of F0 candidate, and selects the final F0 corresponding to the predominant melody as follows:

- For the top-ranked F0 candidate in the current frame, we measure the F0 continuity between the previous and the next frame.

- If the top-ranked F0 candidate of the current frame is not continuous and the F0s of the previous and the next frame are identical, then we take F0 of the previous frame as the final F0 of the current frame.
- If the top-ranked F0 candidate of current frame is not continuous and F0s of the previous and the next frame are not identical, then we measure the continuity of the 2nd-ranked and 3rd-raned F0s of the current frame and take one which is more continuous as the final F0 of the current frame.
- If all of the above conditions are not satisfied, we take the top-ranked F0 candidate of current frame as the starting F0 of a new sound.

### 2.3 Post-Processing

Pitch doubling and halving errors are common problem in pitch extraction, and a post-processing is implemented in order to eliminate these errors. In normal pop music, the F0 of predominant melody is stable in short period and does not deviate from its median F0 by more than ±1.5 octaves. Hence, the proposed post-processing runs as follows:

- Using Eq. (1), we combine F0s into a group $G_l$ with high similarity, where $l$ is a group index, $n$ is a frame index, $f_n$ is F0 of $n$th frame, and $\beta$ is a threshold value. $\beta$ is set to 1.5 tones for allowing small F0 variation. We define the small and large groups based on its cardinality. Then, for the adjacent small and large groups, if the average F0 of small group is equal to the integer multiples of that of large group, the F0 of small group is multiplied by the average F0 ratio between two groups.

$$\begin{array}{ll} if\,(\beta > |f_n - f_{n+1}|) & f_n \in G_l \\ else & l{+}{+} \end{array} \qquad (1)$$

- Next, we re-combine F0s into group $K_l$ using Eq. (1) with $\beta$ = 1 tone. If the average F0 of $K_l$ deviates from the median F0 of all frames by more than ±1.5 octaves, its F0 is corrected by ±1.5 octaves.

### 2.4 Voiced Frame Detection

After the post-processing, we run a voiced frame detection (VAD) process. First, we recalculate the harmonic average energy of each frame using the updated F0 by the post-processing. In addition, we combine the updated F0 into a group using Eq. (1) again and compute the harmonic sharpness of each group. Then we determine the voiced frame using the harmonic average energy and the harmonic sharpness which are suitable to distinguish the vocal from the non-vocal frames.

## 3. PERFORMANCE EVALUATION

We evaluated the performance of the proposed method according to MIREX 2005 presented in [2], and measured the melody detection performance using RPA (raw pitch accuracy) and RCA (raw chroma accuracy). RPA represents the accuracy of the extracted melody frequency based on the ground truth melody frequency and is defined as RPA = (*TPC* + *FNC*)/*GV*, where *GV* represents the number of frames in which vocal melodies are present, *TPC* the number of frames with the correctly extracted frequency out of *GV* frames, and *FNC* the number of the pitch guesses for frames that were judged unvoiced frames. The F0 is considered correct if it is within ±1/4 tone of the ground-truth melody frequency.

Fig. 1 gives the raw pitch accuracy rates using different post-processing methods. Label none corresponds to no post-processing, and RPA rate is 81.47%. When median filter is used, RPA rate is 82.21%, and when the proposed post-processing method is used, RPA rate is increased to 88.22%. After applying VAD to the proposed method, OA (overall accuracy) rate becomes 85.70%.

Fig. 2 shows the pitch contour of ‘opera_fem4’ data. RPA rate is 85.74%, RCA rate is 87.08%, and OA rate is 85.08%.

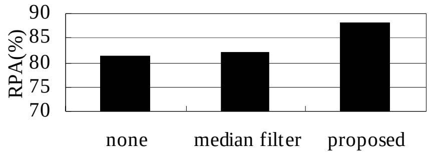

**Figure 1**. RPA results for different post-processing (ADC 2004 DB (all)).

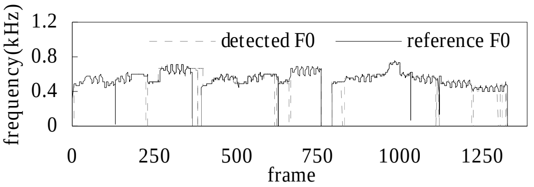

**Figure 2**. Melody extraction results (ADC 2004 DB : opera _ fem4).

## 4. REFERENCES

[1] A. P. Klapuri, “Multiple fundamental frequency estimation by summing harmonic amplitudes,” *Proc. 7th Int. Symposium on Music Information Retrieval*, pp.216-221, Victoria, Canada, Oct. 2006.

[2] Poliner, D. Ellis, A. Ehmann, E. Gomez, S. Streich, B. Ong, “Melody Transcription from Music Audio: Approaches and Evaluation,” *IEEE Trans. Audio, Speech and Language Process.*, vol. 15, no.4, pp.1247-1256, May 2007.