MELODY EXTRACTION BASED ON HARMONIC CODED STRUCTURE

Sihyun Joo **Sanghun Park** **Seokhwan Jo** **Chang D. Yoo**
Department of Electrical Engineering, Korea Advanced Institute of Science and Technology,
373-1 Guseong-dong, Yuseong-gu, Daejeon, 305-701, Korea
{s.joo,psh111,antiland00}@kaist.ac.kr cdyoo@ee.kaist.ac.kr

ABSTRACT

This paper considers a melody extraction algorithm that estimates the melody in polyphonic audio using the harmonic coded structure (HCS) to model melody in the minimum mean-square-error (MMSE) sense. The HCS is harmonically modulated sinusoids with the amplitudes defined by a set of codewords. The considered algorithm performs melody extraction in two steps: i) pitch-candidate estimation and ii) pitch-sequence identification. In the estimation step, pitch candidates are estimated such that the HCS best represents the polyphonic audio in the MMSE sense. In the identification step, a melody line is selected from many possible pitch sequences based on the properties of melody line. Posterior to the melody line selection, a smoothing process is applied to refine spurious pitches and octave errors. The performance of the algorithm is evaluated and compared using the ADC04 and the MIREX05 dataset. The results show that the performance of the proposed algorithm is better than or comparable to other algorithms submitted to MIREX2009.

1. INTRODUCTION

Most people recognize music as a sequence of notes referred to as melody. Melody extraction from polyphonic audio is developed for various applications such as content-based music information retrieval (CB-MIR), audio plagiarism search, automatic melody transcription, music analysis, and query by humming (QBH) [1, 2, 6]. Despite its importance in various applications, melody is not clearly defined [3,4,6]. However, many people consider melody as the most dominant single pitch sequence of a polyphonic audio and the considered algorithm extracts melody following this consideration.

Diverse melody extraction or transcription techniques have been proposed in recent years. Goto introduced a predominant F0 estimation (PreFEst) algorithm [3]. It estimates the weights of prior tone-models over all possible fundamental frequencies(F0s) based on the *maximum a posteriori* (MAP) criterion and determines the F0's temporal continuity by using a multiple-agent architecture. Paiva estimated possible F0s in the short-time Fourier transform (STFT) magnitude domain and decides a single pitch sequence (melody line) based on various properties of melody pitches between near frames [5]. Poliner and Ellis approached the melody line estimation problem as a classification problem and use a support vector machine (SVM) classifier in the estimation [7]. Ryynänen defined an acoustic model based on the hidden-Markov model (HMM) to estimate melody, bass line and chords [1]. Durrieu extracted melody of singing voice by separating singer's voice and background music [2].

There are two main obstacles in extracting accurate melody line [9]. The obstacles are listed below:

1) Accompaniment interference: Accompaniment sound such as harmonics of subdominant melodies and percussive sound acts as noise in the melody pitch estimation.

2) Octave mismatch: Inaccurate melody pitch values which are one octave higher or lower than the ground-truth are often inaccurately estimated: the true melody pitch harmonics appear at either all estimated pitch harmonic locations or every other pitch harmonic locations.

In this paper, an effective melody extraction algorithm that considers the above obstacles is proposed. The algorithm defines a harmonic structure as a model for melody. Related models have been studied for other related applications. Heittola modeled the signal as a sum of spectral bases for sound separation [10]. Duan used pre-coded spectral peak/non-peak position of each possible pitches for pitch tracking [11]. Bay used pre-coded harmonic structure shape for source separation [12]. Goto modeled a pitch harmonics as a Gaussian mixture model [3].

The proposed algorithm minimizes the mean-square error between the given polyphonic audio and the harmonic coded structure (HCS) that is constructed from a codebook

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
© 2011 International Society for Music Information Retrieval.

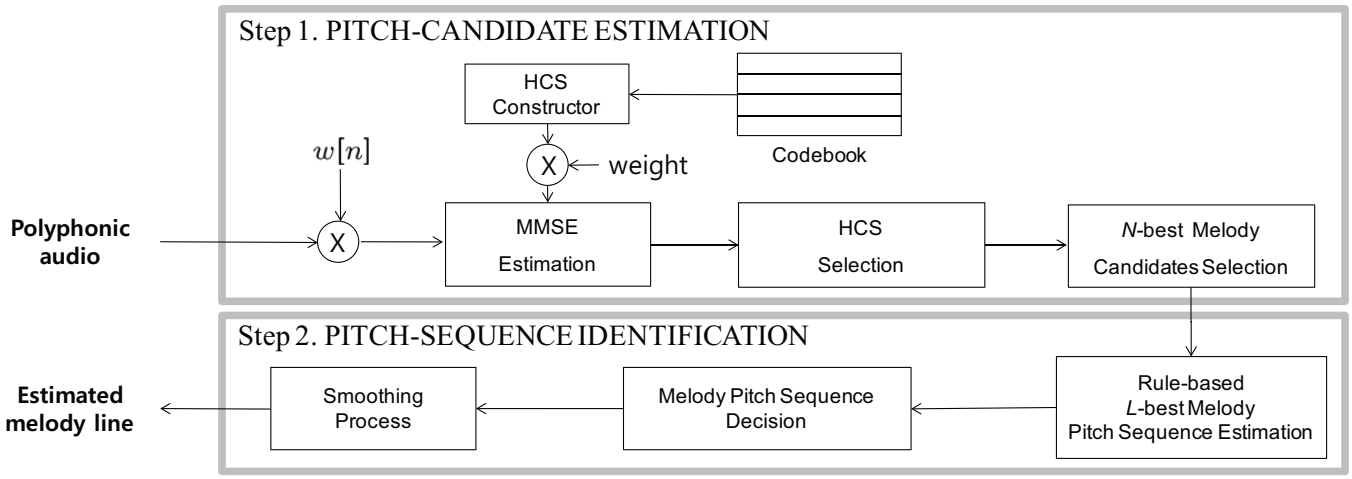


Figure 1. System Overview

of harmonic amplitude set. The codebook was defined by k-means clustering the harmonic amplitudes of training melody data. The algorithm finds N-best pitch candidates for each frame and subsequently determines the best melody line from the pitch candidates by a rule-based identification procedure.

The remainder of this paper is organized as follows. Section 2 describes the proposed melody extraction algorithm. Section 3 shows experimental results of the proposed algorithm and compares the performance to other previous algorithms. Finally, Section 4 concludes this paper.

2. MELODY EXTRACTION ALGORITHM

The overall structure of the proposed algorithm is shown in Figure 1. The proposed algorithm extracts melody pitch sequence (melody line) in two steps: i) pitch-candidate estimation and ii) pitch-sequence identification. In the estimation step, N melody pitch candidates are extracted by finding N most dominant HCS by minimizing minimum-mean-squared error between the magnitude of STFT of framed polyphonic audio using the window function $w[n]$ and a weighted HCS. In the identification step, the melody pitch sequence is estimated based on a certain set of rules of melody line, after which a simple smoothing process is applied. Melody line is decided by first selecting L-best melody line from a sequence of N pitch candidates and then determining the most appropriate melody line from the selection. The smoothing process is performed to remove spurious pitch sequences and octave errors.

2.1 Melody Pitch Candidate Estimation

2.1.1 Construction of HCS

In this paper, a harmonic coded structure (HCS) is proposed to find the dominant melody pitch harmonics in the STFT domain. The windowed harmonic structure can be expressed as follows:

$$h_\eta[n] = w[n] \sum_{m=1}^{H} b_m \cos(m \cdot 2\pi\eta \cdot n + \phi_m),\ H = \lfloor \frac{f_s}{2\eta} \rfloor, \quad (1)$$

where $f_s, \eta, w[n], b_m$, and ϕ_m are sampling frequency, the fundamental frequency (F0) of the HCS, analysis window, amplitude of the mth harmonic, and the phase of the mth harmonic, respectively. The discrete-time Fourier transform (DTFT) of $h_\eta[n]$, $H_\eta(\omega)$, can be expressed as follows:

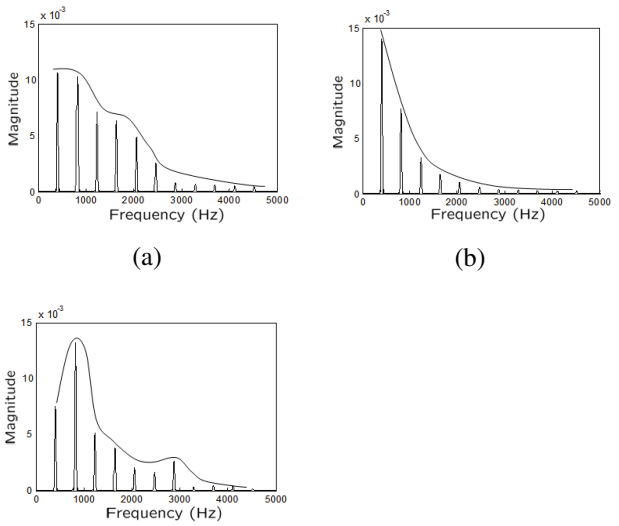
$$H_\eta(\omega) = \sum_{m=1}^{H} B_m W(\omega - m\eta), \quad B_m = b_m e^{-j\phi_m}, \quad (2)$$

where $W(\omega)$ is the DTFT of $w[n]$.

The number of harmonics within a certain bandwidth depends on the pitch and the sampling frequency as defined in (1), but we observe that the harmonic amplitudes tend to decrease with increasing harmonic index ($|B_m| < |B_{m-1}|$ for $m = 2, \cdots, H$). For this reason, we use only 11 harmonics.

The overall envelop of the harmonic amplitudes varies with instrument and pitch [13]. Therefore, it is difficult to construct one fixed melody harmonic structure that fits all the different harmonic amplitude patterns.

To construct a HCS to represent all the different harmonic amplitudes of melody, a codebook is constructed from real audio sample data. Harmonic amplitudes from 26,930 frames of piano sound, 74,631 frames of saxophone sound [14], and 449,430 frames of singing voice [15] are used



(a) (b) (c)

Figure 2. Three estimated harmonic structures when $k = 3$ and the F0 = 400Hz: (a) The first harmonic structure ($i = 1$), (b) the second harmonic structure ($i = 2$), and (c) the third harmonic structure ($i = 3$).

to build the codebook: these three sounds are present as melody in all music considered.

The harmonic amplitude samples are clustered using the extended $k-$means clustering algorithm [16] and the centroids of each cluster are used as codewords. Finally, the HCSs for every possible F0 are constructed using (1) and (2) based on the codebook. Figure 2 illustrates HCSs when $k = 3$ and the F0 = 400Hz.

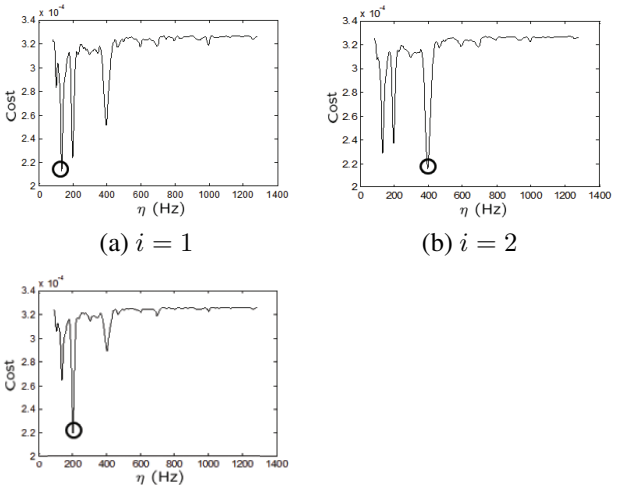
2.1.2 N-Best Melody Pitch Candidates Estimation

The proposed algorithm extracts N melody pitch candidates from each frame of a given polyphonic audio to reduce pitch estimation errors due to accompaniment interference and octave mismatch.

The pitch candidates are estimated based on the consensus that melody is considered as the single dominant pitch sequence in a polyphonic audio. To find the dominant pitch candidates of each frame, a cost function based on the ith HCS, $J_i(\eta, l)$, is defined as follows:

$$J_i(\eta,l) = \int_{-\pi}^{\pi} \Big(|S(\omega,l)| - C_i(\eta,l) \sum_{\substack{m=-H,\\ m\neq 0}}^{H} A_{i,m} |W(\omega - m\eta)| \Big)^2 d\omega, \quad (3)$$

where $S(\omega, l)$ and $C_i(\eta, l)$ are the STFT coefficient of the lth frame at frequency ω and the weight of the ith HCS which is constructed with the ith codeword in the lth frame



(a) $i = 1$ (b) $i = 2$ (c) $i = 3$

Figure 3. The cost of the lth frame given by (3). The circles ($\circ$) indicate $J_i'(l)$ of each HCS.

with F0 = η, respectively. Here, $A_{i,m}$ is the harmonic amplitude of the mth harmonic of the ith codeword. The STFT magnitude of each frame and the HCS with F0 = η satisfy the following constraints:

$$\int_{-\pi}^{\pi} |S(\omega,l)| d\omega = 1, \quad (4)$$

and

$$\int_{-\pi}^{\pi} \sum_{\substack{m=-H,\\ m\neq 0}}^{H} A_{i,m} |W(\omega - m\eta)| d\omega = 1. \quad (5)$$

The HCS represents only the form of the harmonics, not the exact magnitude of harmonics so scaling is required where the weight $C_i(\eta, l)$ is chosen to minimize the cost given in (3), thus

$$\hat{C}_i(\eta,l) = \underset{C_i(\eta,l)}{\operatorname{argmin}} J_i(\eta,l). \quad (6)$$

To find $\hat{C}_i(\eta, l)$, $J_i(\eta, l)$ is differentiated with respect to $C_i(\eta, l)$ and set equal to zero. It yields

$$\hat{C}_i(\eta,l) = \frac{\int_{-\pi}^{\pi} |S(\omega,l)| \Big(\sum_{\substack{m=-H,\\ m\neq 0}}^{H} A_{i,m} |W(\omega - m\eta)| \Big) d\omega}{\int_{-\pi}^{\pi} \Big(\sum_{\substack{m=-H,\\ m\neq 0}}^{H} A_{i,m} |W(\omega - m\eta)| \Big)^2 d\omega}. \quad (7)$$

Prior to extracting melody pitch candidates, the minimum cost of the lth frame using the ith HCS $J_i^{(min)}(l)$ defined below is estimated.

$$J_i^{(min)}(l) = \min_{\eta} \hat{J}_i(\eta,l), \quad (8)$$

where

$$\hat{J}_i(\eta, l) = \int_{-\pi}^{\pi} \Big(|S(\omega, l)| - \hat{C}_i(\eta, l) \sum_{\substack{m=-H, \\ m\neq 0}}^{H} A_{i,m} |W(\omega - m\eta)| \Big)^2 d\omega. \quad (9)$$

Figure 3 shows the cost of each HCS of the lth frame when $k = 3$, and the costs of the circled peaks indicate $J_i^{(min)}(l)$.

Now, the index of the HCS of the lth frame $I(l)$ is estimated by

$$I(l) = \underset{i}{\operatorname{argmin}}\, J_i^{(min)}(l). \quad (10)$$

Generally, harmonic amplitudes of consecutive frames are highly correlated [9]. Thus, the index of HCS that appears frequently within a neighborhood of few frames (including the target frame) should be determined as a more consistent index of the current frame. The updated index of the lth frame is expressed as follows:

$$\hat{I}(l) = \text{mode}[I(l-M), I(l-M+1), \cdots, I(l+M-1), I(l+M)]. \quad (11)$$

where M is the number of neighbor frames considered on either side of the lth frame.

The costs of possible F0s can be finally calculated using (3) with the weight obtained from (7) and the index determined by (11). To obtain a set of N possible melody pitch candidates of the lth frame, the following procedure is performed in obtaining the set $\mathcal{N}_l$ for the lth frame.

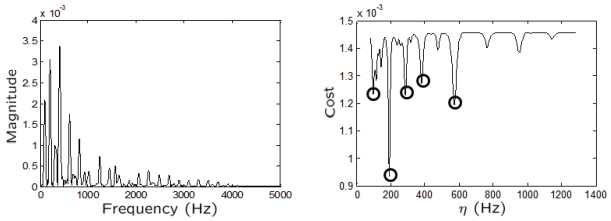
Algorithm 1 N-best Pitch Candidates Determination

$\mathcal{N}_l = \{\}$
for $n = 1, ..., N$ **do**
  $\bar{\eta} = \text{argmin}_{\eta \notin \mathcal{N}_l} J_{\hat{I}(l)}(\eta, l)$
  $\mathcal{N}_l \leftarrow \mathcal{N}_l \cup \bar{\eta}$
end for

Figure 4 (a) and (b) illustrate the STFT magnitude of a frame and its cost, respectively for $N = 5$. The circles in (b) indicate the estimated melody pitch candidates of the frame.

2.2 Melody Pitch Sequence Identification

Once the N-best pitch candidates of each frame are obtained as described in the previous section, a single pitch sequence (melody line) that best represents the melody line is identified. An estimate of the melody line can be obtained by selecting the pitch candidate leading to the minimum cost for each frame. This, however, often leads to inaccurate estimation due to accompaniment interference and octave



(a) (b)

Figure 4. The STFT magnitude and the cost of the lth frame: (a) $|S(\omega, l)|$, (b) the cost of the lth frame obtained by an appropriate HCS.

mismatch. Inaccuracy can be reduced by considering the forward and backward relationship among pitch candidates. The proposed identification algorithm estimates the melody line based on a rule-based method described below.

A more robust melody pitch sequence is obtained by the following two steps: i) *L*-best melody pitch sequences are determined and ii) melody is determined as the melody pitch sequence with the minimum sum cost. (see Figure 1).

2.2.1 L-Best Melody Pitch Sequence Determination

The proposed melody line identification algorithm estimates *L*-best melody lines from *N*-best pitch candidates of each frame based on the following properties of melody line.

P1 The *vibrato* exhibits an extent of $\pm$ 60$\sim$200 *cent* for singing voice and only $\pm$ 20$\sim$30 *cent* for music instruments such as saxophone, violin, and guitar [17].

P2 The note transitions within a musical structure are typically limited to an octave [8].

P3 In general, a rest during singing is longer than 50 ms.

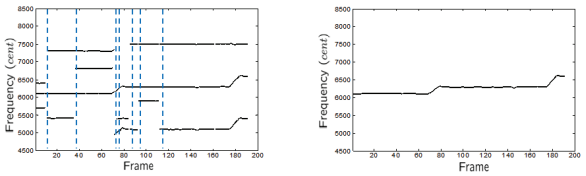
Based on the above properties, the following rules are defined to estimate the melody line.

R1 Any two pitch candidates of successive frames are considered to be included in same melody line segment when the difference between the pitch values is less than the threshold described in **P1**.

R2 When two non-consecutive frames with a time gap less than 50ms have pitch candidates satisfying **P1**, then interpolate between the two pitch values (by **P3**).

R3 When any two pitch candidates of successive frames satisfy only **P2** and not **P1** and **P3**, a transition is assumed to have occurred in the melody line.

In the proposed algorithm, the threshold discussed in **R1** is set to 100 *cent* which was determined experimentally from the validation data. When one of the L-melody lines does not satisfy the given rules, all melody lines are disconnected and a new set of L-melody lines are started.



(a) (b)

Figure 5. Melody pitch sequence estimation: (a) three-best melody pitch sequence estimation, (b) best melody pitch sequence decision.

2.2.2 *Melody Pitch Sequence Decision*

A single melody pitch sequence must be selected from the L-best lines. The best melody pitch sequence is estimated based on the melody definition: melody is a dominant pitch sequence in a polyphonic audio. Hence, after adding up the costs in each melody line segment, the pitch sequence that has the minimum summed-cost is selected as the best melody line segment. Figure 5 (a) and (b) show the result of L-best melody pitch sequence estimation and melody pitch sequence decision, respectively. The vertical dotted lines in (a) represent the disconnecting positions, and the pitch sequences between two vertical dotted lines are considered as melody line candidates.

2.2.3 *Smoothing Process*

Although the procedures described in Section 2.2.1 and 2.2.2 effectively reduce accompaniment interference and octave mismatch, it is difficult to estimate the true melody pitch sequence if the interference occurs throughout the melody line. Thus, a smoothing process is applied to find a more robust melody line.

After the single melody pitch sequence is estimated, spurious sequences are removed and replaced with interpolated pitch values between non-spurious pitches. The spurious sequence is determined by following conditions. i) A pitch sequence which switches to another note and returns to the original note within short time is considered as the spurious sequence. ii) A pitch sequence which has a transition over one octave is also regarded as an inaccurate estimate.

3. EVALUATION

Two CD-quality (16-bit quantization, 44.1 kHz sample rate) test datasets are used for evaluation. One dataset used for the evaluation is the Audio Description Contest (ADC) 2004 dataset, and the other is the Music Information Retrieval Evaluation eXchange (MIREX) 2005 dataset. Table 1 shows the configurations of the evaluation datasets.

In the experiment, the possible fundamental frequency range is set from 80Hz (3950 *cent*) to 1280Hz (8750 *cent*) and 3 clusters are used for building codebook ($k = 3$). In the melody pitch candidate estimation step, 3-best pitch candidates are chosen for each frame ($N = 3$) and the number of neighbor frames for deciding harmonic structure is set to 7 ($M = 7$). In the melody pitch sequence identificaion step, 3-best melody lines are estimated ($L = 3$). These values are determined experimentally.

Dataset	Melody	Number of files
ADC04	Vocal melody	8
	Nonvocal melody	12
MIREX05	Vocal melody	9
	Nonvocal melody	4

Table 1. Evaluation dataset.

Dataset	Algorithms	RPA (%)	RCA (%)
ADC04	Cao et al.	85.1	86.3
	Durrieu et al.	81.4	83.4
	Hsu et al.	63.9	73.6
	Dressler	87.1	87.6
	Wendelboe	82.3	86.4
	Cancela	82.9	83.4
	Rao et al.	76.9	85.1
	Tachibana et al.	61.0	71.8
	Proposed	**81.8**	**86.0**
MIREX05	Ryynänen et al. [1]	67.3	69.1
	Durrieu et al. [19]	74.5	79.6
	Tachibana et al. [20]	74.0	76.7
	Proposed	**76.1**	**80.7**

Table 2. Result Comparison.

The estimated melody pitch is considered correct when the absolute value of the difference between the ground-truth and the estimated pitch frequency is less than quarter tone (50 *cent*). This is shown as

$$|F_g(l) - F_e(l)| \leq \frac{1}{4} tone\ (50cent), \qquad (12)$$

where $F_g(l)$ and $F_e(l)$ denote ground-truth and estimated pitch frequency of the lth frame, respectively.

The performance of the proposed algorithm is evaluated with row pitch accuracy (RPA) and row chroma accuracy (RCA) [8].

Table 2 shows the evaluation results for all algorithms considered. The results on the ADC04 dataset are from the MIREX 2009 homepage [18]. When obtaining the results on the MIREX05 dataset, we referred the results in [20] or used the codes publicly released by the authors [1, 21]. The best result on each dataset is underlined, and the result of the proposed algorithm is highlighted in bold. The proposed

algorithm achieved the best performance both in RPA and RCA on the MIREX05 dataset. It also performed comparably to the other algorithms on the ADC04 dataset.

4. CONCLUSION

In this paper, an algorithm extracting melody from a polyphonic audio using the HCS which is constructed from the codebook of harmonic amplitude set obtained by k-means clustering is considered. The algorithm focuses on reducing accompaniment interference and octave mismatch. The algorithm consists of two steps: N-best pitch candidates estimation step and rule-based melody identification step. First, multiple pitch candidates of each frame are estimated using the cost function which determines the most dominant HCS of the frame in the MMSE sense. Second, a single pitch sequence (melody line) is identified based on certain rules of melody line. To handle the spurious pitch sequence problem, the smoothing process is applied. The considered algorithm is tested on two datasets: the ADC04 dataset and the MIREX05 dataset. Experimental results show that the proposed algorithm is better than or comparable to the other melody extraction algorithms.

5. REFERENCES

[1] M. P. Ryynänen and A. P. Klapuri: "Automatic transcription of melody, bass line, and chords in polyphonic music," *Computer Music Journal*, Vol.32, No.3, pp. 72–86, 2008.

[2] J.-L. Durrieu, G. Richard, and B. David: "Singer melody extraction in polyphonic signals using source separation methods," *in Proceedings of the ICASSP*, 2008.

[3] M. Goto: "A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communcation*, Vol.43, No.4, pp. 311–329, 2004.

[4] R. P. Paiva: "An approach for melody extraction from polyphonic audio: Using perceptual principles and melodic smoothness," *The Journal of the Acoustical Society of America*, Vol.122, No.5, pp. 2962–2969, 2007.

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

[6] V. Rao and P. Rao: "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music," *IEEE ASLP*, Vol.18, No.8, pp. 2145–2154, 2010.

[7] G. E. Poliner and D. P. W. Ellis: "A classification approach to melody transcription," *in Proceedings of the ISMIR*, 2005.

[8] G. E. Poliner, D. P. W. Ellis, and A. F. Ehmann: "Melody transcription from music audio: approach and evaluation," *IEEE ASLP*, Vol.15, No.4, pp. 1247–1256, 2007.

[9] S. Jo, and C. D. Yoo: "Melody extraction from polyphonic audio based on particle filter," *in Proceedings of the ISMIR*, 2010.

[10] T. Heittola, A. Klapuri and T. Virtanen: "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," *in Proceedings of the ISMIR*, 2009.

[11] Z. Duan, J. Han and B. Pardo: "Harmonically informed multi-pitch tracking," *in Proceedings of the ISMIR*, 2009.

[12] Mert Bay, James W. Beauchamp: "Harmonic source separation using prestored spectra," *in Proceedings of the ICA*, pp. 561–568, 2006.

[13] Beauchamp, J. W.: "Analysis and Synthesis of Musical Instrument Sounds" *in Analysis, Synthesis, and Perception of Musical Sounds: The Sound of Music*, ed (Springer), pp. 1–89, 2007.

[14] L. Fritts: *University of Iowa musical instument samples*, http://theremin.music.uiowa.edu/MIS.html.

[15] C.-L. Hsu and J.-S. R. Jang: *MIR-1K Dataset*, http://sites.google.com/site/unvoicedsoundseparation/mir-1k.

[16] D. Pelleg and A. W. Moore: "X-means: Extending K-means with efficient estimation of the number of clusters," *in Proceedings of the ICML*, 2000.

[17] R. Timmers and P. W. M Desain: "Vibrato: the questions and answers from musicians and science," *the International Conference on Music Perception and Cognition*, 2000.

[18] *MIREX2009:Audio Melody Extraction Results*, http://www.music-ir.org/mirex/wiki/2009:Audio_Melody_Extraction_Results.

[19] J.-L. Durrieu, G. Richard, B. David, and C. Févotte: "Source/Filter Model for Unsupervised Main Melody Extraction From Polyphonic Audio Signals" *IEEE ASLP*, Vol.18, No.3, pp. 564–575, 2010.

[20] H. Tachibana, T. Ono, N. Ono, and S. Sagayama: "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source," *Proceedings of the ICASSP*, 2010.

[21] http://www.durrieu.ch/phd/software.html