MELODY EXTRACTION FROM POLYPHONIC AUDIO BASED ON PARTICLE FILTER

Seokhwan Jo **Chang D. Yoo**
Department of Electrical Engineering, Korea Advanced Institute of Science Technology,
373-1 Guseong-dong, Yuseong-gu, Daejeon, 305-701, Korea
antiland00@kaist.ac.kr cdyoo@ee.kaist.ac.kr

ABSTRACT

This paper considers a particle filter based algorithm to extract melody from a polyphonic audio in the short-time Fourier transforms (STFT) domain. The extraction is focused on overcoming the difficulties due to harmonic / percussive sound interferences, possibility of octave mismatch, and dynamic variation in melody. The main idea of the algorithm is to consider probabilistic relations between melody and polyphonic audio. Melody is assumed to follow a Markov process, and the framed segments of polyphonic audio are assumed to be conditionally independent given the parameters that represent the melody. The melody parameters are estimated using sequential importance sampling (SIS) which is a conventional particle filter method. In this paper, the likelihood and state transition are defined to overcome the aforementioned difficulties. The SIS algorithm relies on sequential importance density, and this density is designed using multiple pitches which are estimated by a simple multi-pitch extraction algorithm. Experimental results show that the considered algorithm outperforms other famous melody extraction algorithms in terms of the raw pitch accuracy (RPA) and the raw chroma accuracy (RCA).

1. INTRODUCTION

Many people believe that people recognize music as a sequence of monophonic notes called melody, and for this reason, melody extraction is playing an important role in music content processing which has recently become an important research area. Although the debate over the definition of melody is on going [1–3], many experts concur that melody should be the dominant pitch sequence of a polyphonic audio. In this paper, melody is defined to be the singing voice pitch sequence in the vocal part and the pitch sequence of the solo instrument in non-vocal part or non-vocal music. When a music contains singing voice, most people recognize music by the vocal melody line in the vocal part. However, in non-vocal part such as intermezzo and non-vocal music such as jazz and orchestra, most people recognize music by the melody line of the solo instrument.

Many melody extraction algorithms have been proposed over the last one decade [1–6], albeit with limited success. Melody extraction from the polyphonic audio is still difficult for the following reasons:

1. Harmonic interference: Harmonics of other instrument signal interfere in the estimation of the melody pitch harmonics.

2. Percussive sound interference: Percussive sound interfere to estimate the melody pitch because the energy of it forms a vertical ridge with strong and wide-band spectral envelopes.

3. Octave mismatch: The estimated pitch can be one octave higher or lower than the ground-truth.

4. Dynamic variation in melody: Accurate pitch estimation in the beginning, end and sudden transient regions of a melody is difficult.

In this paper, melody pitch frequency and harmonic amplitudes that represent the melody are estimated in the short-time Fourier transforms (STFT) domain. The main idea of the algorithm is to consider a probabilistic relations between melody and polyphonic audio. Melody pitch frequency and harmonic amplitudes are assumed to follow Markov processes, and the framed segments of polyphonic audio are assumed to be conditionally independent given melody pitch frequency and harmonic amplitudes. Thus, melody pitch frequency and harmonic amplitudes can be estimated from the polyphonic audio based on the Bayesian sequential model once the likelihood and state transition are defined. The likelihood is defined to be robust to harmonic and percussive sound interferences. The state transition of melody pitch frequency is adjusted by control parameters that discourages octave mismatch and dynamic variation in the melody. The sequential importance sampling (SIS) algorithm, a conventional particle filter algorithm, is used to estimate the melody parameters. The SIS algorithm relies on a so-called sequential importance density, and this density is designed using multiple pitches which are estimated by a simple multi-pitch extraction algorithm.

This paper is organized as follows. Section 2 presents the melody extraction from polyphonic audio based on particle filter. Section 3 provides experimental results. Finally, Section 4 concludes this paper.

2. MELODY EXTRACTION FROM POLYPHONIC AUDIO BASED ON PARTICLE FILTER

2.1 Melody extraction from polyphonic audio

The melody pitch harmonics $x_t[n]$ in the tth frame is defined as follows:

$$x_t[n] = w[n] \sum_{m=1}^{H} A_{m,t} \cos(m\omega_{0,t} n + \phi_{m,t}), \quad (1)$$

where $A_{m,t}$, $\omega_{0,t}$, $\phi_{m,t}$, H and $w[n]$ are the amplitude of the mth harmonic in the tth frame, the melody pitch frequency in the tth frame, the phase of the mth harmonic in the tth frame, number of melody pitch harmonics, and the analysis window function, respectively. The polyphonic audio can be expressed as

$$z_t[n] = x_t[n] + y_t[n], \quad (2)$$

where $z_t[n]$ and $y_t[n]$ are the polyphonic audio signal and signal of other instruments in the tth frame, respectively. In the frequency domain, the following relationship holds:

$$\mathbf{z}_t = \mathbf{x}_t + \mathbf{y}_t, \quad (3)$$

where $\mathbf{z}_t$, $\mathbf{x}_t$, and $\mathbf{y}_t$ are the N-point discrete Fourier transforms (DFT) of $z_t[n]$, $x_t[n]$, and $y_t[n]$, respectively.

The parameters of the melody pitch harmonics – the melody pitch frequency and the harmonic amplitudes – must be estimated for the melody extraction. This paper assumes that the phase of the melody pitch harmonics is the same as the phase of the polyphonic audio, i.e., the phase of the melody pitch is not estimated since human ear is assumed to be unsensitive to phase variations. Thus, the tth frame parameter set is defined as

$$\mathbf{\Theta}_t = (\omega_{0,t}, \mathbf{A}_t), \quad (4)$$

where $\mathbf{A}_t = [A_{1,t}, A_{2,t}, ..., A_{H,t}]$. The objective of melody extraction is to estimate $\mathbf{\Theta}_t$ from given $\mathbf{z}_t$. It is usually observed that successive parameters – $\omega_{0,t}$ and $\mathbf{A}_t$ – are highly correlated. In this paper, it is assumed that $\mathbf{\Theta}_t$ is considered a Markov process and $\mathbf{y}_t$ at each frame is conditionally independent given $\mathbf{\Theta}_t$. Here, $\mathbf{\Theta}_t$ is considered latent while $\mathbf{y}_t$ is observed. From this perspective, the Bayesian sequential model for melody extraction can be constructed as shown in Figure 1. In Figure 1, $p(\mathbf{z}_t|\mathbf{\Theta}_t)$, $p(\mathbf{\Theta}_t|\mathbf{\Theta}_{t-1})$, and ρ_t are likelihood, state transition, and control parameter to decide the state transition of the melody pitch frequency, respectively. From this Bayesian sequential model, the posterior probability $p(\mathbf{\Theta}_{0:t}|\mathbf{z}_{1:t})$ [1] is estimated, and it is used to estimate $\mathbf{\Theta}_t$ for melody extraction. To estimate $p(\mathbf{\Theta}_{0:t}|\mathbf{z}_{1:t})$, likelihood and state evolution equations with state transition needs to be defined.

[1] The notation $a_{0:t}$ means that $a_{0:t} = [a_0, a_1, ..., a_t]^T$

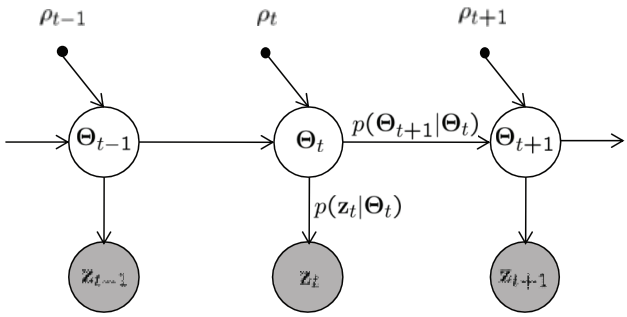


Figure 1. Bayesian sequential model for melody extraction. $\mathbf{z}_t$, $\mathbf{\Theta}_t$, and ρ_t are polyphonic audio, melody parameter ($\omega_{0,t}$ and $\mathbf{A}_t$), and control parameter, respectively.

To obtain the likelihood, it is assumed that the DFT coefficients of $\mathbf{y}_t$ follow a zero mean complex multivariate Gaussian distribution, which is given by

$$\mathbf{y}_t \sim \mathcal{N}(0, \Sigma_t), \quad \Sigma_t = \mathrm{diag}(\sigma_{t,1}^2, \sigma_{t,2}^2, ..., \sigma_{t,N}^2), \quad (5)$$

where Σ_t and $\sigma_{t,k}$ are the covariance matrix in tth frame and the variance of the kth bin in the tth frame, respectively. Eqn. (5) yields the likelihood as follows:

$$\begin{aligned} p(\mathbf{z}_t|\mathbf{\Theta}_t) &= \mathcal{N}(\mathbf{z}_t; \mathbf{x}_t, \Sigma_t) \\ &\propto \exp\left\{-(\mathbf{z}_t - \mathbf{x}_t)^H \Sigma_t^{-1} (\mathbf{z}_t - \mathbf{x}_t)\right\}, \end{aligned} \quad (6)$$

where $(\cdot)^H$ is the Hermitian operator. To define $p(\mathbf{z}_t|\mathbf{\Theta}_t)$, $\sigma_{t,k}$ must be estimated. In this paper, $\sigma_{t,k}$ is estimated using the decision-directed method [7] as follows:

$$\widehat{\sigma}_{t,k} = \alpha \widehat{\sigma}_{t-1,k} + (1-\alpha)|Y_{t,k}|^2, \quad (7)$$

where α and $Y_{t,k}$ are a smoothing factor and the kth bin DFT coefficient of $\mathbf{y}_t$, respectively. However, Eqn. (7) can not be used directly since $Y_{t,k}$ is unknown. It is assumed that $Y_{t,k}$ is highly correlated with $Y_{t-1,k}$. Therefore, the estimation is modified as follows:

$$\widehat{\sigma}_{t,k} = \alpha \widehat{\sigma}_{t-2,k} + (1-\alpha)|\widehat{Y}_{t-1,k}|^2. \quad (8)$$

Accurate estimation of Σ_t will lead to robustness to harmonic and percussive sound interferences. Figure 2 shows an example of $\mathbf{z}_t$ and an estimate of Σ_t, and it is easily shown that the likelihood in Eqn. (6) is maximized at the true $\mathbf{\Theta}_t$.

The state evolution equations, which describe relationships of the parameters at frame t, are set as follows:

$$A_{m,t} = A_{m,t-1} + v_{A,t-1}, \quad (9)$$

$$\omega_{0,t} = \omega_{0,t-1} + v_{\omega_0,t-1}, \quad (10)$$

where $v_{A,t-1}$ and $v_{\omega_0,t-1}$ are the random perturbations corresponding to harmonic amplitudes and melody pitch frequency of the $(t-1)$th frame, respectively. This type of state evolution equations is called *random walk*: the current state is a random perturbation of the previous state. It is important to define $p(v_{A,t-1})$ and $p(v_{\omega_0,t-1})$ accurately, and in this paper, $p(v_{A,t-1})$ is assumed to be a truncated Gaussian as shown in Figure 3 since $A_{m,t} > 0$, and

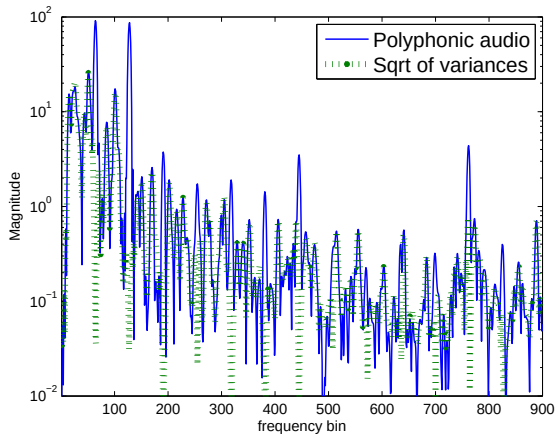


Figure 2. Example of polyphonic audio ($\mathbf{z}_t$) and the estimated variances (Σ_t) of other instrument signal.

$p(v_{\omega_0,t-1})$ is assumed to be a Gaussian whose variance controlled by ρ_t. Melody line is characterized by prolonged periods of smoothness, with infrequent sharp changes in note transition or during vibrato regions.

Furthermore, there are two general rules concerning the melody line: 1) the vibrato exhibits an extent of 60∼200 cents [2] for singing voice and only 20∼30 cents for other [8], and 2) the transitions are typically limited to one octave [1]. Therefore, assumption that $v_{\omega_0,t-1}$ follows a Gaussian distribution with fixed variance is not appropriate. In this paper, the state transition from the from the $(t-1)$th state to the tth state of the melody pitch frequency is controlled by ρ_t which indicates the degree of the melody line being whether in transition or not. Here, transition includes vibrato. And, ρ_t is defined as

$$\rho_t = \widehat{\omega}_{0,t-1} - \widehat{\omega}_{0,t-2}, \tag{11}$$

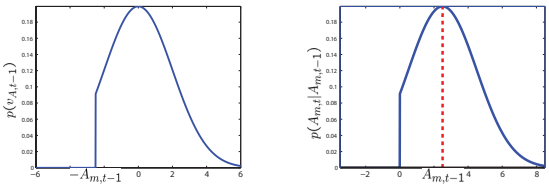
and $p(v_{\omega_0,t-1})$ is given by

$$p(v_{\omega_0,t-1}) = \begin{cases} \mathcal{N}(0, 20 \text{ cent}) & \rho_t < 50 \text{ cent} \\ \mathcal{N}(0, 50 \text{ cent}) & 50 \text{ cent} \le \rho_t < 100 \text{ cent} \\ \mathcal{N}(0, 100 \text{ cent}) & 100 \text{ cent} \le \rho_t \end{cases}. \tag{12}$$

When ρ_t is small, the current melody pitch frequency represents a certain note frequency and has a value similar to the previous melody pitch frequency. When ρ_t is large, the current melody pitch frequency is with high probability in a note transition or vibrato regions and has a value dissimilar to the previous melody pitch frequency. The state transition of melody pitch frequency defined by Eqn. (12) can lead to robustness to octave mismatch and dynamic variation in melody.

[2] The *cent* is a unit of logarithmic frequency range, and it is defined as

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



(a) $p(v_{A,t-1})$ (b) $p(A_{m,t}|A_{m,t-1})$

Figure 3. State transition in harmonic amplitudes.

2.2 Melody extraction based on particle filter

In this paper, $p(\mathbf{\Theta}_{0:t}|\mathbf{z}_{1:t})$ is approximated using Monte Carlo integration and $\mathbf{\Theta}_t$ is estimated using the particle filter. The SIS algorithm which is a common particle filter method [9, 10] is adopted to estimate the parameters of the melody. If the likelihood and the state transition follow a Gaussian distribution, the problem can be solved by Kalman filter. However, the state transition is not assumed to be a Gaussian. The SIS algorithm is used to obtain $p(\mathbf{\Theta}_{0:t}|\mathbf{z}_{1:t})$ based on the Bayesian sequential model shown as Figure 1.

The posterior density $p(\mathbf{\Theta}_{0:t}|\mathbf{z}_{1:t})$ can be approximated as follows:

$$p(\mathbf{\Theta}_{0:t}|\mathbf{z}_{1:t}) \approx \sum_{i=1}^{N_p} w_t^{(i)} \delta(\mathbf{\Theta}_{0:t} - \mathbf{\Theta}_{0:t}^{(i)}), \tag{13}$$

where $\mathbf{\Theta}_{0:t}^{(i)}$, $w_t(i)$, and N_p are the ith particle of $\mathbf{\Theta}_{0:t}$, associated weight, and the number of particles, respectively. The weights are normalized such that $\sum_{i=1}^{N_p} w_t^{(i)} = 1$. The weights are chosen using the method of importance sampling. If the particle $\mathbf{\Theta}_{0:t}^{(i)}$ were drawn from an importance density $q(\mathbf{\Theta}_{0:t}^{(i)}|\mathbf{z}_{1:t})$, the weights in Eqn. (13) are defined as follows:

$$w_t^{(i)} \propto \frac{p(\mathbf{\Theta}_{0:t}^{(i)}|\mathbf{z}_{1:t})}{q(\mathbf{\Theta}_{0:t}^{(i)}|\mathbf{z}_{1:t})}. \tag{14}$$

If the importance density is chosen to factorize as follows

$$q(\mathbf{\Theta}_{0:t}|\mathbf{z}_{1:t}) = q(\mathbf{\Theta}_t|\mathbf{\Theta}_{0:t-1}, \mathbf{z}_{1:t}) q(\mathbf{\Theta}_{0:t-1}|\mathbf{z}_{1:t-1}), \tag{15}$$

then one can obtain particles $\mathbf{\Theta}_{0:t}^{(i)} \sim q(\mathbf{\Theta}_{0:t}^{(i)}|\mathbf{z}_{1:t})$ by augmenting each of the existing particles $\mathbf{\Theta}_{0:t-1}^{(i)} \sim q(\mathbf{\Theta}_{0:t-1}^{(i)}|\mathbf{z}_{1:t-1})$ with the new state $\mathbf{\Theta}_t^{(i)} \sim q(\mathbf{\Theta}_t|\mathbf{\Theta}_{0:t-1}, \mathbf{z}_{1:t})$. The weight update equation can be derived as follows using Eqn. (14) and Eqn. (15)

$$w_t^{(i)} \propto w_{t-1}^{(i)} \frac{p(\mathbf{z}_t|\mathbf{\Theta}_t^{(i)}) p(\mathbf{\Theta}_t^{(i)}|\mathbf{\Theta}_{t-1}^{(i)})}{q(\mathbf{\Theta}_t^{(i)}|\mathbf{\Theta}_{t-1}^{(i)}, \mathbf{z}_t)}. \tag{16}$$

A common problem with the particle filter is the degeneracy phenomenon, where after a few iterations, most particles have negligible weight [9,10]. A suitable measure of degeneracy is the effective particle size, N_{eff}, which is given by

$$\widehat{N_{eff}} = \frac{1}{\sum_{i=1}^{N_p} (w_t^{(i)})^2}. \tag{17}$$

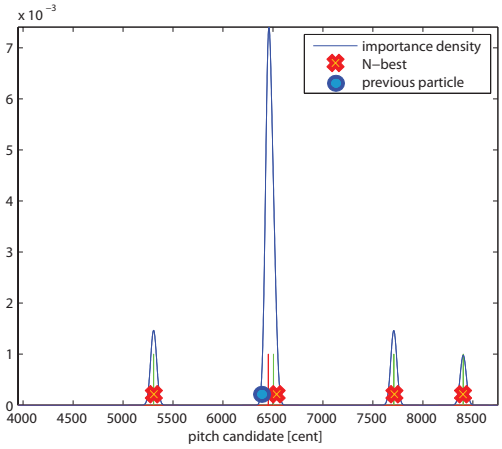


Figure 4. Design of $q(\omega_{0,t}^{(i)}|\omega_{0,t-1}^{(i)}, \mathbf{z}_t)$.

In this paper, to avoid the degeneracy problem, resampling algorithm is used when $N_{eff} \leq \frac{N_p}{2}$.

Finally, estimation of parameters is achieved by posterior mean after obtaining $p(\mathbf{\Theta}_{0:t}|\mathbf{z}_{1:t})$.

$$\widehat{\omega}_{0,0:t} = \sum_{i=1}^{N_p} w_t^{(i)} \omega_{0,0:t}^{(i)}, \tag{18}$$

$$\widehat{\mathbf{A}}_{0:t} = \sum_{i=1}^{N_p} w_t^{(i)} \mathbf{A}_{0:t}^{(i)}. \tag{19}$$

2.2.1 *Design of sequential importance density*

The performance of the SIS algorithm depends on the choice of $q(\mathbf{\Theta}_t^{(i)}|\mathbf{\Theta}_{t-1}^{(i)}, \mathbf{z}_t)$. Setting $q(\mathbf{\Theta}_t^{(i)}|\mathbf{\Theta}_{t-1}^{(i)}, \mathbf{z}_t) = p(\mathbf{\Theta}_t^{(i)}|\mathbf{\Theta}_{t-1}^{(i)})$ leads to not only unnecessary large number of particles but also difficulties in estimating $p(\mathbf{\Theta}_{0:t}|\mathbf{z}_{1:t})$. In this paper, a multiple pitch estimation algorithm is used to define $q(\mathbf{\Theta}_t^{(i)}|\mathbf{\Theta}_{t-1}^{(i)}, \mathbf{z}_t)$ since the melody pitch frequency is assumed to be one of the pitch estimate given by the multiple pitch estimates. A main idea in defining $q(\mathbf{\Theta}_t^{(i)}|\mathbf{\Theta}_{t-1}^{(i)}, \mathbf{z}_t)$ is to generate particles of the melody parameters similar to the estimated multiple pitch parameters. To obtain multiple pitch parameters, the multiple pitch estimation algorithm proposed in [11] is used.

Before drawing particles from the importance density, $q(\mathbf{\Theta}_t^{(i)}|\mathbf{\Theta}_{t-1}^{(i)}, \mathbf{z}_t)$ is factorized as follows:

$$\begin{aligned} & q(\omega_{0,t}^{(i)}, \mathbf{A}_t^{(i)}|\omega_{0,t-1}^{(i)}, \mathbf{A}_{t-1}^{(i)}, \mathbf{z}_t) \\ = \; & q(\mathbf{A}_t^{(i)}|\omega_{0,t}^{(i)}, \mathbf{A}_{t-1}^{(i)}, \mathbf{z}_t) q(\omega_{0,t}^{(i)}|\omega_{0,t-1}^{(i)}, \mathbf{z}_t). \end{aligned} \tag{20}$$

Here, $\omega_{0,t}$ and $\mathbf{A}_t$ are considered conditionally independent given $\omega_{0,t-1}^{(i)}$, $\mathbf{A}_{t-1}^{(i)}$, and $\mathbf{z}_t$. First, melody pitch particles are drawn as given by

$$\omega_{0,t}^{(i)} \sim q(\omega_{0,t}^{(i)}|\omega_{0,t-1}^{(i)}, \mathbf{z}_t), \tag{21}$$

where $q(\omega_{0,t}^{(i)}|\omega_{0,t-1}^{(i)}, \mathbf{z}_t)$ is shown as Figure 4. In defining $q(\omega_{0,t}^{(i)}|\omega_{0,t-1}^{(i)}, \mathbf{z}_t)$, the current melody pitch particles are drawn near the N-best pitch candidates obtained from the multiple-pitch estimation and the melody pitch particles drawn in the previous frame. After drawing melody pitch particles, melody pitch harmonic amplitudes particles are drawn as given by

$$\begin{aligned} & \mathbf{A}_t^{(i)} \sim q(\mathbf{A}_t^{(i)}|\omega_{0,t}^{(i)}, \mathbf{A}_{t-1}^{(i)}, \mathbf{z}_t) \\ = \; & \mathcal{N}\left(\frac{\mathbf{A}_{t-1}^{(i)} + \mathbf{A}_t^{\mathbf{z}^{\omega_{0,t}^{(i)}}}}{2}, \frac{|\mathbf{A}_{t-1}^{(i)} - \mathbf{A}_t^{\mathbf{z}^{\omega_{0,t}^{(i)}}}|}{2} \right) \end{aligned} \tag{22}$$

where $\mathbf{A}_t^{\mathbf{z}^{\omega_{0,t}^{(i)}}}$ is the harmonic amplitudes corresponding pitch candidate near $\omega_{0,t}^{(i)}$ with constraint $\mathbf{A}_t^{(i)} > 0$. In defining $q(\mathbf{A}_t^{(i)}|\omega_{0,t}^{(i)}, \mathbf{A}_{t-1}^{(i)}, \mathbf{z}_t)$, the current harmonic amplitude particles which are similar to the previous harmonic amplitude particles and harmonic amplitudes of the N-best pitch candidates are generated. If $\mathbf{A}_{t-1}^{(i)}$ and $\mathbf{A}_t^{\mathbf{z}^{\omega_{0,t}^{(i)}}}$ are similar, then $\frac{|\mathbf{A}_{t-1}^{(i)} - \mathbf{A}_t^{\mathbf{z}^{\omega_{0,t}^{(i)}}}|}{2} \approx 0$, therefore, $\mathbf{A}_t^{(i)} \approx \frac{\mathbf{A}_{t-1}^{(i)} + \mathbf{A}_t^{\mathbf{z}^{\omega_{0,t}^{(i)}}}}{2}$. If $\mathbf{A}_{t-1}^{(i)}$ and $\mathbf{A}_t^{\mathbf{z}^{\omega_{0,t}^{(i)}}}$ are not similar, then $\frac{|\mathbf{A}_{t-1}^{(i)} - \mathbf{A}_t^{\mathbf{z}^{\omega_{0,t}^{(i)}}}|}{2} >> 0$, therefore, $\mathbf{A}_t^{(i)}$ is generated somewhat randomly.

The outline of the considered algorithm is given below.

Outline of the considered algorithm

Melody extraction based on the SIS

For $i = 1, ..., N_p$

1. Generate the particles
   - Melody pitch particles $\omega_{0,t}^{(i)} \sim q(\omega_{0,t}^{(i)}|\omega_{0,t-1}^{(i)}, \mathbf{z}_t)$
   - Harmonic amplitudes particles $\mathbf{A}_t^{(i)} \sim q(\mathbf{A}_t^{(i)}|\omega_{0,t}^{(i)}, \mathbf{A}_{t-1}^{(i)}, \mathbf{z}_t)$
2. *Update the weights*: Eqn. (16)

Normalize the weights ($\sum_{i=1}^{N_p} w_t^{(i)} = 1$).

Resampling: Resampling algorithm is used when $N_{eff} \leq \frac{N_p}{2}$.

Estimation: Melody pitch frequency in tth frame is estimated by Eqn. (18). Harmonic amplitudes of melody pitch harmonics in tth frame are estimated by Eqn. (19).

---

3. EVALUATION

The considered algorithm was evaluated and compared to other melody extraction algorithms using the ISMIR 2004 Audio Description Contest (ADC04) database. The database contains 20 polyphonic musical audio pieces. All test data are single channel PCM data with 44.1 kHz sample rate and 16-bit quantization. Table 1 shows the data composition of the ADC04 set. Search range of melody pitch frequency was between 80Hz and 1280Hz in frequency do-

Melody Instrument	Sytle
Synthesized voice (4)	POP
Saxophone (4)	Jazz
MIDI instruments (4)	Folk(2), Pop(2)
Human voice (2 male, 2 female)	Classical opera
Male Voice (4)	POP

Table 1. Summary of ADC04 data set. The number in parentheses is the number of corresponding pieces.

	RPA	RCA
Goto [2]	65.8% (2005)	71.8% (2005)
Paiva el al. [3]	62.7% (2005)	66.7% (2005)
Marlot [4]	60.1% (2005)	67.1% (2005)
Ryynanen el al. [5]	68.6% (2005)	74.1% (2005)
Ellis el al. [6]	73.2% (2006)	76.4% (2006)
Considered algorithm	77.3%	83.8%

Table 2. Result comparison. The number in parentheses is the year when their algorithms were submitted to the MIREX.

main (3950 cent and 8750 cent in cent domain). The Hanning window was used with 48ms frame length and 10ms frame hop size. $\alpha = 0.98$ in Eqn. (8) was used. $N_p = 500$ in Eqn. (13) was used.

The estimated melody is correct when the absolute value of the difference between the ground-truth frequency and estimated frequency is less than 50 cent ($\frac{1}{4}$ tone). The performance of the considered algorithm was evaluated in terms of raw pitch accuracy (RPA) and raw chroma accuracy (RCA). The RPA is defined as the proportion of frames in which the estimated melody pitch is within $\pm\frac{1}{4}$ tone of the reference pitch. And the RCA is defined in the same manner as the raw pitch accuracy; however, both the estimated and reference frequencies are mapped into a single octave in order to forgive octave transpositions.

The considered algorithm was compared to the other famous melody extraction algorithms such as algorithms proposed by Goto [2], Paiva et al. [3], Marlot [4], Ryynanen el al. [5], and Ellis et al. [6]. Their performances are based on results of the Music Information Retrieval Evaluation eXchange (MIREX) [12].

Table 2 shows the evaluation results for all algorithms considered. The considered algorithm outperformed the others in terms of the RPA and the RCA. The difference between the RPA and RCA is proportional octave mismatch error. Although the algorithm in this paper is considered to be robust against octave mismatch, the difference between the RPA and the RCA is 6.5 %. The multiple pitch estimation algorithm proposed in [11] was quite simple and vulnerable to octave error, i.e., inaccuracy in sequential importance density led to inaccurate melody pitch candidates.

4. CONCLUSION

The melody extraction algorithm from the polyphonic audio based on particle filter is considered in this paper. Most people recognize music as not all of note sequences but a special monophonic note sequence called melody. However, melody extraction from polyphonic audio is difficult due to the following impediments: harmonic interference, percussive sound interference, octave mismatch, and dynamic variation in melody. The main idea of the algorithm is to consider probabilistic relations between melody and polyphonic audio. Melody is assumed to follow a Markov process, and the framed segments of polyphonic audio are assumed to be conditionally independent given the parameters that represent the melody. The parameters are estimated using the SIS algorithm. This paper shows that likelihood and state transition that are required in the SIS algorithm are defined to be robust against the aforementioned impediments. The performance of the SIS algorithm depends on a sequential importance density, and this density is designed by multiple pitch. Experimental results show that the considered 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] 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, No. 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 signals: exploiting perceptual rules, note salience, and melodic smoothness," *Computer Music Journal*, Vol. 30, No. 4, pp. 80–98, 2006.

[4] M. Marolt: "On finding melodic lines in audio recordings," *Proceeding of 7th International Conference on Digital Audio Effects DAFx 04*, pp. 217–221, 2004.

[5] M. P. Ryynanen and A. P. Klapuri: "Note event modeling for audio melody extraction," *MIREX 2005 Audio Melody Extraction Contest*, 2005.

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

[7] Yariv Ephraim: "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 32, No. 6, pp. 1109–1121, 1984.

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

[9] A. Doucet, N. de Freitas, and N. J. Gordon: *Sequential Monte Carlo methods in practice*, Springer-Verlag, New York, 2001.

[10] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp,: "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, Vol. 50, No. 2, pp. 174–188, 2002.

[11] S. Joo, S. Jo, and C. D. Yoo: "Melody extraction from polyphonic audio signal MIREX 2009," *MIREX 2009 Audio Melody Extraction Contest*, 2009.

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