EXTENDED ABSTRACT FOR AUDIO MELODY EXTRACTION IN MIREX 2010 (TO BE MODIFIED)

Hideyuki Tachibana, Takuma Ono, Nobutaka Ono and Shigeki Sagayama
Graduate School of Information Science and Technology, The University of Tokyo
Hongo 7-3-1, Bunkyo, Tokyo, Japan
{tachibana, tono, onono, sagayama}@hil.t.u-tokyo.ac.jp

ABSTRACT

This extended abstract is for the "Audio Melody Extraction" evaluation in MIREX2010. We describe an algorithm that estimates the melody line from a music audio signal. The algorithm is comprised of two stages: melodic component enhancement and melody line tracking. Only a few researchers used this approach because of difficulties of the melody enhancement. Our enhancement algorithm focuses on temporal variability of melodic source, e.g., vibrato, spectral fluctuation, of singing voice, violin, etc. After enhancement, we estimate the melody line by a simple tracking algorithm. The method is going to be evaluated in MIREX 2010.

1. INTRODUCTION

Melodies are the most attractive parts of music for most listners. For this reason, technologies which relate to melodies, e.g., automataic karaoke generation, melody transcription, etc., may attract interests from music fans and professional musicians. Therefore, development of melody extraction techniques has much significance as a fundamental techniques for those applications. Though it is not difficult for humans to recognize melodies from accompaniments, it is a very challenging task for computers. Some of the difficulties of automatic melody transcription are caused by the similarities between melodies and accompaniments, i.e., both accord with the same chords, rhythms.

In consequence, melody extraction from music audio signals has been an important issue to study recently, and Audio Melody Extraction (AME) evaluation has been held in Music Information Retriecal EXchange (MIREX). Many researchers have participated in the AME evaluation for the last few years, and the evaluation is going to be also held in year 2010 [1].

Our submission to AME MIREX2010 is basically based on the paper [2]. The method focuses on temporal-variability of melodics source: quasi-periodic fluctuation of F_0 and amplitude (e.g., vibrato of singing voice, violin, etc.,) transience and instantaneous onset of melodic notes compared to sustained chords, and spectral fluctuation of the melodic sounds. Using those features of melodic component, we first enhance the component by a filtering algorithm which is called Multi-stage HPSS. Then, we apply a simple tracking algorithm for monophionic music audio signals. The sequential approach has been employed only by a few researchers because of difficulties of melody enhancement.

In the enhancement stage, we focus on temporal variability of melodic source. The temporally-variable components can be enhanced by multi-staged harmonic/percussive sound separation (Multi-Stage-HPSS), a particular filtering algorithm. The aim of the stage is to suppress the accompanimental components which interfere with the subsequent tracking process.

The tracking stage is formulated as a maximum a posteriori (MAP) estimation problem. The objective function of MAP estimation is the sum of a transition score defined between a time frame and the succeeding time frame and a state score defined as likelihood function of F_0 in each frame. The optimal solution to the problem can be obtained effectively by dynamic programming which binds locally-optimal solutions into the globally-optimal solution.

2. MELODIC COMPONENT ENHANCEMENT

2.1 Harmonic/Percussive Sound Separation (HPSS)

We first introduce a fundamental signal processing algorithm, called Harmonic/Percussive Sound Separation (HPSS) [3–5]. The algorithm originally is a method to separate a music audio signal into "harmonic components" and "percussive components." Despite the name of the method, HPSS utilizes neither harmonic structures of sound nor the prior knowledge of percussions. Instead, the method uses only information of "smoothness" of the sounds: harmonic sounds are "smooth" in time direction, and percussive sounds are "smooth" in frequency direction, because the former are stationary and periodic for a short period of time, whereas the latter are transient and aperiodic.

2.2 Temporal Variability of Melodic Component

Some musical sources such as singing voice and unfletted strings sometimes contain fluctuation. Beside, melodic notes do not sustain for a long time. The former can be considered as the broadness of bandwidth, and the latter, as

the shortness of duration. Therefore, if we set some parameters properly in HPSS calculations, we can make HPSS treat those temporal-variable components as "percussions" though they are not apparently percussion and HPSS with ordinary parameters treat those components as "harmonic." Actually, it depends on the time-frequency resolution of spectrogram, i.e., the length of windows functions of short-time Fourier transform (STFT) calculation.

2.3 Multi-stage HPSS

To sum up the previous section, HPSS can separate a same signal in two different ways as described below:

1. Separate the music audio signal into "sustained (chord) sound" and "temporally-variable (melody) sound + instantaneous (percussive) sound" by HPSS on long-framed STFT domain (approximately 100–500[ms]).

2. Separate the music audio signal into "sustained (chord) sound + temporally-variable (melody) sound" and "instantaneous (percussive) sound" by HPSS on short-framed STFT domain (approximately 15–50[ms]).

Consequently, by combining those two processings, we can enhance melodic components in a music audio signal.

$$\mathrm{Input}(t) \xrightarrow{\text{HPSS with long window}} \{H^{(1)}(t), P^{(1)}(t)\}, \quad (1)$$

$$P^{(1)}(t) \xrightarrow{\text{HPSS with short window}} \{H^{(2)}(t), P^{(2)}(t)\}. \quad (2)$$

The obtained $H^{(2)}(t)$ is the desired melodic-component-enhanced signal. We call the two-stage processing as multi-stage HPSS [2].

3. PITCH TRACKING AND VOICING DETECTION

3.1 Pitch Tracking

Given a spectrogram of melodic-component-enhanced signal S_n, we consider the way to search the melody line X_n that maximize the probability $p(S_n, X_n)$:

$$\ln p(S_\tau, X_\tau) = \ln p(s_\tau|x_\tau) + \ln p(x_\tau|x_{\tau-1}) + \ln p(S_{\tau-1}, X_{\tau-1}), \quad (3)$$

where s_τ is a short-time constant Q [6] spectrum of the observed melodic-component-enhanced signal, and x_τ is the hidden state: pitch of the melody which is to be estimated in the problem. S_τ and X_τ are $S_\tau = \{s_1, \ldots, s_\tau\}, X_\tau = \{x_1, \ldots, x_\tau\}$ respectively.

We model the likelihood function $p(s_\tau|x_\tau)$ by matched filtering between s_τ and timbre model on log-frequency domain. We assumed n-th harmonics of the timbre has $1/n$ amplitude of fundamental frequency.

We model the probability function density of melody transition $p(x_\tau|x_{\tau-1})$ as Gaussian function:

$$\ln p(x_\tau|x_{\tau-1}) = -\frac{1}{2\sigma^2}(x_\tau - x_{\tau-1})^2, \quad (4)$$

because large leaps of melody occur only occationaly.

3.2 Voicing Detection (DRAFT)

This section is the novelty of our submission to MIREX2010 compared to MIREX2009. To discriminate voiced and non-voiced frames, we used $H^{(2)}(t)$ and $P^{(2)}(t)$. Let $\{v_\tau\}_{0\le\tau\le n}$ and $\{p_\tau\}_{0\le\tau\le n}$ be the short-time power of $H^{(2)}(t)$ and $P^{(2)}(t)$, where τ denotes the frame indices.

As $H^{(2)}(t)$ consists of melodic components and some noises, it is conceivable that applying some discrimination algorithm to $\{v_\tau\}$ can discriminate voiced frame from non-voiced frame. However, it is also a difficult problem to estimate noise statistics of the $H^{(2)}(t)$ because there still remain such problems that we mentioned in the introduction in $H^{(2)}(t)$ a little. Hence, we supposedly assumed that the noise statistics of $H^{(2)}(t)$ can be approximated by the statistics of $\{p_\tau\}$,

Based on such an assumption, we compared Mahalanobis distance of each $v_\tau^{0.3}$ from $E[v_\tau^{0.3}]$ and $E[p_\tau^{0.3}]$, to determine which each frame τ should be estimated as voiced or non-voiced.

4. MIREX2010 EVALUATION

The method is going to be evaluated in MIREX2010 [1].
(DRAFT)

5. CONCLUSION

In this extended abstract, we described a melody extraction algorithm. The algorithm comprises melodic component enhancement and pitch tracking. The enhancement algorithm focuses on temporal-variability of melodic source, and separate them by HPSS on two differently resoluted spectrograms.
(DRAFT)

6. REFERENCES

[1] `http://www.music-ir.org/mirex/wiki/2010:Audio_Melody_Extraction`

[2] H. Tachibana, T. Ono, N. Ono, S. Sagayama: "Melody Line Estimation in Homophonic Music Audio Signals Based on Temporal-Variability of Melodic Source," *Proc. ICASSP*, pp.425-428, Mar., 2010.

[3] N. Ono, K. Miyamoto, H. Kameoka, J. Le Roux, Y. Uchiyama, E. Tsunoo, T. Nishimoto, and S. Sagayama, "Harmonic and Percussive Sound Separation and its Application to MIR-related Tasks," *Advances in Music Information Retrieval, ser. Studies in Computational Intelligence, Z. W. Ras and A. Wieczorkowska, Eds. Springer,* 274, pp.213-236, Feb., 2010.

[4] N. Ono, K. Miyamoto, H. Kameoka, S. Sagayama: "A Real-Time Equalizer of Harmonic and Percussive Components in Music Signals," *Proceedings of the International Symposium on Music Information Retrieval,* pp. 139–144, 2008.

[5] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, S. Sagayama: "Separation of a Monuaral Audio Signals into Harmonic/Percussive Components by Complementary Diffusion on Spectrogram," *Proceedings of EUSIPCO*, 2008.

[6] J. C. Brown, "An effecient algorithm for the calculation of a constant Q transform," *Journal of Acoustic Society of America*, Vol.92, No. 5, pp.2698–2701, 1992.