MELODY EXTRACTION IN MUSIC AUDIO SIGNALS BY MELODIC COMPONENT ENHANCEMENT AND PITCH TRACKING

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" contest of MIREX2009. 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 vaiability of melodic source, e.g., vibrato of singing voice, violin, etc. After enhancement, we estimate the melody line by a simple tracking algorithm. The method is evaluated in MIREX2009, and it is confirmed that the method is effective if the melody is played by singing voice, especially in low SNR conditions.

1. INTRODUCTION

Melodies are the most attractive parts of music for most listners. For this reason, melody-related technologies, 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 recognition are caused by the similarities between melodies and accompaniments. E.g., both accord with the same chords, rhythms.

This paper focuses on temporal-variability of melodics source: quasi-periodic fluctuation of F_0 and amplitude (e.g., vibrato of singing voice, violin, etc.,) and transience and instantaneous onset of melodic notes compared to sustained chords. Using those features of melodic component, we first enhance the component by a filtering algorithm which was developed by us [1, 2]. Then, we apply a simple tracking algorithm for monophionic music audio signals [3]. The sequential approach has been employed by only 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 [1, 2]. 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 transition score defined between a time frame and the following time frame and state score defined as most likely F_0 estimation 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) [4, 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 of F_0 and amplitude. Beside, melodic notes do not sustain for a long time. In a physical point of view, 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 + temporally-variable (melody) sound” and “instantaneous (percussive) sound” by HPSS on SHORT-framed STFT domain (approximately 15–50[ms]).

2. 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]).

Consequently, by combining those two processings, we can enhance melodic components in a music audio signal. The two-stage processing we call Multi-Stage HPSS [1,2].

3. PITCH TRACKING

Given a spectrogram S_n, we consider the way to search the melody line X_n that maximize the following probability $p(S_n, X_n)$:

$$\ln p(S_t, X_t) = \ln p(s_t|x_t) + \ln p(x_t|x_{t-1}) + \ln p(S_{t-1}, X_{t-1}), \quad (1)$$

where s_t is a short-time constant Q [6] spectrum of the observed melodic-component-enhanced signal, and x_t is the hidden state: pitch of the melody which is to be estimated in the problem. S_t and X_t are $S_t = \{s_1, \dots, s_t\}, X_t = \{x_1, \dots, x_t\}$ respectively.

We model the likelihood function $p(s_t|x_t)$ by matched filtering between s_t 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 $p(x_t|x_{t-1})$ as Gaussian function:

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

because large leaps of melody occur only occationaly.

4. MIREX2009 EVALUATION

The method was evaluated in MIREX2009. The evaluation was conducted using several datasets under several conditions. The benchmarks ware Voicing Detection, Voicing False Alarm, Raw Pitch Accuracy, Raw Chroma Accuracy and Overall Acuuracy. As our method does not discriminate voiced/unvoiced segments, Voicing Detection, Voicing Falm Alarm, and Overall Accuracy are not significant, but Raw Pitch Accuracy and Raw Chroma Accuracy are principal concern here.

We show the excerpted results about MIREX09 dataset, which consists of 374 pieces, melodies of which are played by singing voice. Table 1 shows results on MIREX09 dataset under -5dB conditions, and Table 2 shows results on the same dataset under 0dB conditions. In those cases, our method marked the highest Raw Pitch Accuracy and Raw Chroma Accuracy in 12 algorithms. The results verify the effectiveness of our melodic component enhancement algorithm.

Table 3 shows results on the same dataset under $+5$dB conditions. Our algorithm marked a relatively high performance also in this case, though not as good as in low SNR cases. Other detailed results are available in [7].

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. By evaluations in MIREX2009, it is verified that our algorithm is effective especially in low SNR conditions.

Our future works include improvement of pitch tracking algorithm for monophonic music audio signals, embedding voiced/unvoiced recognition model into pitch tracking algorithm, and use of “melodic-component-suppressed signal” which can be obtained in the process of the enhancement.

6. REFERENCES

[1] H. Tachibana, N. Ono, S. Sagayama: “Enhancement and Suppression of Vocal Components in Music Audio Signals Based on Temporal Variability of Spectra,” *IPSJ Technical Report*, MUS-81, No.12, 2009 (in Japanese)

[2] H. Tachibana, N. Ono, S. Sagayama: “Vocal Sound Suppression in Monaural Audio Signals by Multi-stage Harmonic-Percussive Sound Separation (HPSS),” *Proceedings of ASJ Spting Meeting*, 2-8-8, pp. 853-854, 2009 (in Japanese)

[3] H. Tachibana, T. Ono, N. Ono, S. Sagayama: “Melody Line Estimation in Music Audio Signals Based on Spectral Fluctuation of Singing Voice,” *Proceedings of ASJ autumn meeting*, 2009 (in Japanese)

[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.

[7] `http://www.music-ir.org/mirex/2009/index.php/Audio_Melody_Extraction_Results`

Table 1. MIREX 2009 Audio Melody Extraction Summary results – MIREX 2009 Dataset – -5dB mix. Excerpted 5 participants and 3 benchmarks.

	Raw Pitch Accuracy	Raw Chroma Accuracy	Overall Accuracy
Tachibana, Ono, Ono, Sagayama	74.8896%	78.5338%	48.6449%
Dressler	62.4877%	66.2816%	51.6864%
Joo, Jo, Yoo	58.5304%	64.7866%	42.2335%
Rao, Rao	54.6785%	58.7592%	43.3962%
Durrieu, Richard, David (1)	53.7796%	58.0902%	45.5482%

Table 2. MIREX 2009 Audio Melody Extraction Summary results – MIREX 2009 Dataset – 0dB mix. Excerpted 5 participants and 3 benchmarks.

	Raw Pitch Accuracy	Raw Chroma Accuracy	Overall Accuracy
Tachibana, Ono, Ono, Sagayama	82.2943%	85.7474%	53.5623%
Dressler	80.4565%	81.8811%	68.2237%
Joo, Jo, Yoo	75.9354%	80.2461%	49.686%
Hsu, Jang, Chen (1)	72.6577%	75.2906%	53.1752%
Durrieu, Richard, David (1)	69.8804%	72.5138%	60.1294%

Table 3. MIREX 2009 Audio Melody Extraction Summary results – MIREX 2009 Dataset – $+5$dB mix. Excerpted 5 participants and 3 benchmarks.

	Raw Pitch Accuracy	Raw Chroma Accuracy	Overall Accuracy
Dressler	89.1898%	89.6585%	78.4061%
Hsu, Jang, Chen (1)	84.8561%	86.5939%	74.9723%
Tachibana, Ono, Ono, Sagayama	84.8473%	88.289%	55.6746%
Joo, Jo, Yoo	84.3853%	87.6795%	51.7425%
Durrieu, Richard, David (1)	80.8947%	82.2161%	72.7971%