MELODY LINE ESTIMATION IN HOMOPHONIC MUSIC AUDIO SIGNALS BASED ON TEMPORAL-VARIABILITY OF MELODIC SOURCE

Hideyuki Tachibana, Takuma Ono, Nobutaka Ono, Shigeki Sagayama

Graduate School of Information Science and Technology, The University of Tokyo
Hongo 7-3-1, Bunkyo, Tokyo, Japan

ABSTRACT

Estimation of melody line in homophonic music audio signals is a challenging subject of study. Some of the difficulties are derived from presence of accompanying components. To overcome those difficulties, we propose a method to enhance melodic components in music audio signals. The enhancement algorithm uses fluctuation and shortness of melodic components, which we call temporal-variability. We also discuss a melody tracking algorithm, which can be simple thanks to the preprocessing. In this paper, we describe the enhancement method and tracking method, and show the experimental results that supports the efficiency of our methods.

Index Terms— melodic component enhancement, melody transcription, harmonic-percussive sound separation, temporal-variability

1. INTRODUCTION

Most musical pieces which we daily listen to are homophonic music, which comprise a leading melody and accompanying chords and percussions. It is not a very difficult problem for humans to find the melody line in a homophonic music signals. In contrast, it is a very hard problem for computers to extract the melody line from the music audio signal. There are several factors which makes it difficult. An example is the presence of accompanying chord parts and rhythm parts. As there are many shared natures between melody and those accompanying sounds (e.g., both melodies and chords have harmonic structures, and both melodies and rhythms are played rhythmically), it is difficult to discriminate melodies from other sounds. Thus it is naturally expected that suppressing accompanying components might improve melody estimation. However, it has also been a difficult problem because those accompanying sounds do not comply with some natures of "noise" which have been supposed in other fields: whiteness, stationarity, and noncorrelatedness to the signal.

In accord with those conditions, several melody detectors in homophonic music have been proposed. Goto's PreFEst[1] estimates predominant melody line and bass line simultaneously, by maximum a posteriori (MAP) estimation. Cao et al. focused on harmonic structure of melody, and extracted melody lines by harmonic tracking [2]. Additionally, a competition on Audio Melody Extraction has been held recently as a part of MIREX [3] (a MIR-related contest), and many participants have submitted their algorithms to the contest.

Meanwhile, we confront the problem the way we mentioned above, i.e., first, we enhance melodic components (suppress accompanying components), and subsequently, we track the most likely melody line. Such two stage approaches have been taken only a few times until today (e.g., [4]), because of difficulties of the first stage. In melodic-component-enhancement, to avoid the difficulties, we focus on fluctuation of F_0 and amplitude (sometimes called vibrato), and shortness of duration, of melodic tones. In this paper, we describe the melodic-component-enhancer and subsequent melody tracker, and show an experimental result that supports our method.

2. ENHANCEMENT OF MELODIC COMPONENTS

2.1. Harmonic/Percussive Sound Separation

Before getting into the main subject of the section, we describe a brief overview of Harmonic/Percussive Sound Separation (HPSS) [5, 6], which is developed in our laboratory. HPSS is an algorithm to separate a music signal into "harmonic components" and "percussive components." Despite the name of the method, HPSS does not depend on harmonic structures of sound, nor prior knowledge of percussions. Instead, the method uses only information of "anisotropic smoothness" of the sounds. Specifically, we define the "anisotropic smoothness" of sound as partial differentials of the spectrogram in temporal or frequency direction: harmonic components are "smooth in temporal direction" because they are sustained and periodic for a while; percussive components are "smooth in frequency direction" because they are instantaneous and aperiodic.

HPSS exploits anisotropic smoothness of "harmonic sound" and "percussive sound" to separate them. Specifically, HPSS is designed as an optimization problem to minimize

such an objective functional as

$$J[H,P] = w_{\mathrm{H}} \iint \left(\frac{\partial}{\partial t} |H(t,\omega)|^{\gamma} \right)^2 dt d\omega + w_{\mathrm{P}} \iint \left(\frac{\partial}{\partial \omega} |P(t,\omega)|^{\gamma} \right)^2 dt d\omega \quad (1)$$

under a constraint that sum of $H(t,\omega)$ and $P(t,\omega)$ should be as close to the original signal as possible: a simplest instantiation is

$$H(t,\omega) + P(t,\omega) = W(t,\omega), \quad (2)$$

where $H(t,\omega)$ and $P(t,\omega)$ are complex spectrograms of "harmonic sound" and "percussive sound" to be estimated, $W(t,\omega)$ is the spectrogram of the original signal, w_{H} and w_{P} are constants which uniform the physical dimension of each term, and γ is an exponential constant approximately 0.6 (to imitate the auditory systems). The concept of each term of the Eq. (1) is as follows: the first term lays a restriction on $H(t,\omega)$ to be "smooth" in temporal direction, and the second term lays a restriction on $P(t,\omega)$ to be "smooth" in frequency direction. Instantiations of those concepts alternative to Eq. (1) and (2) are also possible, some of which are described in the original papers [5, 6].

In practice, we interpret the equations in discrete form, and solve them numerically. An instantiation is successive over relaxation (SOR)-like updating formulae, derived from the condition that the solution should satisfy in extremum of Eq. (1). Let $w_{\mathrm{H}} = w_{\mathrm{P}} = 1$ (this means they should be equivalent to temporal resolution and frequency resolution respectively in discrete form) and $\gamma = 0.5$ to simplify the calcutation, and we can derive following updating formulae:

$$\lambda_{t,\omega} = |H_{t+1,\omega}|^{0.5} + |H_{t-1,\omega}|^{0.5}, \quad (3)$$

$$\mu_{t,\omega} = |P_{t,\omega+1}|^{0.5} + |P_{t,\omega-1}|^{0.5}, \quad (4)$$

$$|H_{t,\omega}| \leftarrow \frac{\lambda_{t,\omega}^2 |W_{t,\omega}|}{\lambda_{t,\omega}^2 + \mu_{t,\omega}^2}, \; |P_{t,\omega}| \leftarrow \frac{\mu_{t,\omega}^2 |W_{t,\omega}|}{\lambda_{t,\omega}^2 + \mu_{t,\omega}^2}, \quad (5)$$

where t and ω represent indices of discrete time and frequency. Note that the spectrogram $W_{t,\omega}$ should be calculated by an inversible transformation because we need audible waveforms later; short-time Fourier transform (STFT) with half-over-wrapping square-root hanning window (sine window) suffices the condition.

2.2. Temporal-Variability of Melodic Tones

Some of melodic tones contain 4–8 Hz quasi-periodic vibrations of F_0-s, some of which are called vibrato. Additionally, melodies are not sustained for a long while (e.g., each note of a melody does not last for 2 or 3 bars except extreme cases), but they change in a little while, typically in fourth or eighth note. Conversely, chord tones which are played by such instruments as guitar and piano etc., do not fluctuate very much, and they are maintained stationary for a while, typically the length of half to several bars, except some cases, e.g., arpeggio. We call those natures of melodies as "temporal-variability" apart from chord tones' "temporal-stability."

2.3. Effects of Window Function of Spectrogram and Procedure of Melodic Source Enhancement

The time-frequency resolution of a spectrogram cannot be infinite, because of the uncertainty principle of the spectral analysis, but it is finite and the resolution depends on the length of the window function of STFT. Our concernment in this section is to regulate the temporally-variable components' action into which to be separated H or P in HPSS calculation, by adjusting the time-frequency resolution of STFT.

If we calculate STFT with long window function ($\approx$ 200[ms]), the spectrogram has low temporal resolution and high frequency resolution. This biased resolution makes sounds whose length is long and bandwidth is narrow (e.g., temporally stable sounds) appear "smoothly" in temporal direction (H); sounds whose length is moderately-or-very short, and bandwidth is moderately-or-very broad (e.g., percussive sounds and temporally-variable sounds) appear "smoothly" in frequency direction (P).

In contrast, if we use short window function ($\approx$ 30[ms]), the result of the separation is quite different. In this case, the spectrogram has high temporal resolution and low frequency resolution. This biased resolution makes sounds whose length is moderately-or-very long, and bandwidth is moderately-or-very narrow (e.g., temporally-variable and temporally-stable sounds) appear "smoothly" in temporal direction (H); sounds whose length is short and bandwidth is broad (e.g., percussive sounds) appear "smoothly" in frequency direction (P).

Therefore, we can regulate which temporally-variable sounds to be separated into by adjusting the length of window function. In summary, we can enhance temporally-variable sounds by following two-stage HPSS:

$$W(t) \xrightarrow{\text{HPSS with long window}} \{H^{(1)}(t), P^{(1)}(t)\}, \quad (6)$$

$$P^{(1)}(t) \xrightarrow{\text{HPSS with short window}} \{H^{(2)}(t), P^{(2)}(t)\}. \quad (7)$$

The obtained $H^{(2)}(t)$ is the desired melody-enhanced signal.

3. MELODY TRACKING BY DYNAMIC PROGRAMMING

3.1. Framework of Tracking Algorithm

We estimate the melody line in the melodic-components-enhanced signal obtained in the previous section using a probabilistic framework based on suppositions that the pitch of a melody should be close to the adjacent one, and the sound of a melody should have a harmonic structure.

Let S_n be a time series of short-time constant Q [7] spectra $S_n = \{\mathrm{s}_t\}_{1 \le t \le n}$, where $\mathrm{s}_t = \{s_{t,x}\}_{1 \le x \le m}$, and $X_n =$

$\{x_t\}_{1\leq t\leq n}$ be the time series of the instantaneous pitch x_t of the melody which is to be estimated. Given S_n, we estimate the most likely melody line $\hat{X}_n$ which maximize the a posteriori probability function $p(X_n|S_n)$, i.e.,

$$\hat{X}_n = \underset{X_n}{\operatorname{argmax}}\, p(X_n|S_n) = \underset{X_n}{\operatorname{argmax}}\, p(X_n, S_n). \quad (8)$$

Here, if we assume that a pitch x_t depends only on the last pitch x_{t-1}, and a spectrum s_t depends only on the simultaneous pitch x_t, we can rewrite Eq. (8) by following recurring equation:

$$\ln p(X_t, S_t) = \ln p(X_{t-1}, S_{t-1}) + \ln p(\mathbf{s}_t|x_t) + \ln p(x_t|x_{t-1}). \quad (9)$$

We can treat the increments of the recurrent formula frame by frame. Therefore, the optimization problem is divided into local optimization in each frame separately, and the globally optimal solution to the problem can be obtained effectively by dynamic programming.

3.2. Pitch Likelihood Model $p(\mathbf{s}_t|x_t)$ and Pitch Transition Model $p(x_t|x_{t-1})$

To obtain the pitch likelihood $p(\mathbf{s}_t|x_t)$ in each frame, we applied matched filtering. The filter is intended to suppresses the harmonics (or, enhances the F_0 of harmonic components). The filter can be executed as correlation operation between the short-time constant Q spectrum s_t and a sound model:

$$\hat{s}_t(x_t) = \sum_{\xi=0}^{\xi_{\mathrm{Nyq}}-x_t} s_t(x_t+\xi)q(\xi), \quad (10)$$

where x_t and ξ are discrete log frequency, ξ_{Nyq} is Nyquist frequency, and $q(\xi)$ is the sound model, whose n-th harmonic has $1/n$ amplitude of the F_0. The obtained $\hat{s}_t(x_t)$ should have large value if x_t was the true F_0 of this frame, then we directly assumed $\hat{s}_t(x_t)$ to be the likelihood function $p(\mathbf{s}_t|x_t)$.

For pitch transition model, we assumed transition probability between a pitch x_{t-1} to the subsequent pitch x_t as Gaussian function, as melodies do not jump large intervals all of a sudden, but the nearer pitch to the last pitch x_{t-1} is likely to be the subsequent pitch:

$$p(x_t|x_{t-1}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_t - x_{t-1})^2}{2\sigma^2}\right\}. \quad (11)$$

4. EXPERIMENTAL EVALUATION

4.1. Experimental Conditions

We conducted an experiment on melody extraction. We exploited a referential dataset of MIREX provided by LabROSA at Columbia University (available at [8]). The dataset contains 13 audio files and ground truth F_0 data for each audio file. The audio files are CD-quality (PCM, 16-bit, 44.1 kHz), monaural, 20–30[s] length short clips. 9 of 13 clips are vocal songs, and other 4 clips are instrumental pieces generated by MIDI. In the ground truth data, F_0 of the melody is given every 10[ms]. As we do not need very high harmonics, we resampled all clips into 16 kHz.

As criteria, we used Raw Pitch Accuracy (RPA) and Raw Chroma Accuracy (RCA), both of them are used in MIREX. RPA is the ratio of correct frames to all melody-active frames. A frame is regarded "correct" if difference between the estimated pitch and the ground truth of the frame are within a quarter tone (50 cents). RCA is similarly calculated, but the octave errors are ignored.

In the enhancement stage, i.e., two-stage HPSS, the lengths of window functions were set 256[ms] (4096 samples) and 32[ms] (512 samples). We executed HPSS by applying Eq. (3), (4), and (5) 30 times for all bins, keeping phase spectrogram unchanged: $\angle H_{t,\omega} = \angle P_{t,\omega} = \angle W_{t,\omega}$.

In the tracking stage, we used following parameters. The search range of x_t was between 165Hz and 660 Hz. The frequency resolution was $1/10$ semitone (10 cents). Q value of constant Q transform was 60.0 (approximately equivalent to quarter semitone). Time resolution Δt of constant Q transform was 10[ms]. We set $\sigma^2 = 0.2$ [semitones^2/ms], practically, $\sigma^2\Delta t$ was used.

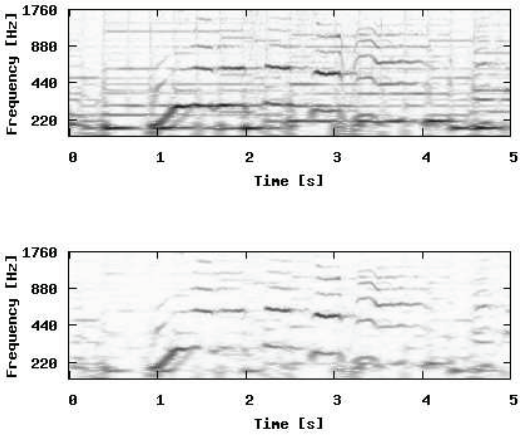

Fig. 1. Spectrograms of original signal (excerpted 5[s] from train06.wav, LabROSA dataset) and melodic-component-enhanced signal.

4.2. Results

Fig. 1 shows an effect of two-stage HPSS. The upper is spectrogram of the original signal, and the lower is that of melodic-component-enhanced signal. Horizontal lines and vertical lines, viz., sustained sounds and percussive sounds, seen in the original spectrogram became faint after the en-

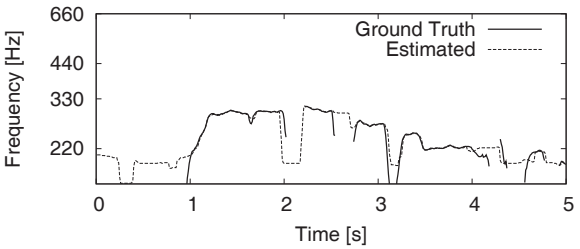

Fig. 2. Estimated melody line and ground truth of train06.wav (excerpted 5[s]).

Table 1. Raw Pitch Accuracy and Raw Chroma Accuracy of melody tracking for LabROSA dataset.

	Enhancement & tracking		Tracking only (for reference)	
	RPA	RCA	RPA	RCA
train 01	92.2%	92.2%	94.4%	94.4%
train 02	76.1%	77.4%	66.4%	69.4%
train 03	60.0%	69.7%	58.1%	58.1%
train 04	78.0%	83.2%	70.1%	70.5%
train 05	92.2%	92.2%	91.0%	91.0%
train 06	69.7%	71.7%	53.6%	53.6%
train 07	82.1%	82.2%	77.4%	78.5%
train 08	88.0%	88.0%	83.1%	84.7%
train 09	84.6%	87.7%	76.8%	76.8%
train 10	99.7%	99.7%	93.8%	93.8%
train 11	69.9%	69.9%	73.8%	73.8%
train 12	25.6%	35.2%	13.0%	21.7%
train 13	43.4%	48.4%	13.3%	30.8%

hancement, while melodic components are maintained. Fig. 2 shows a result of the pitch tracking. The estimated line was approximately equal to the ground truth, except melody-absent parts (0.0 – 0.9[s], 2.0 – 2.2[s], etc.)

Left half of Table 1 shows the accuracy ratios for all data. There were not large differences between RPA and RCA; it indicated that the octave errors (false detection of the harmonics) did not occur very often. Comparing to the result of a referential experiment conducted without enhancement stage (right half of Table 1), it was shown that enhancement stage basically improved the accuracy. Performances on train 01 to 09 were generally high, as the melodies were played by singing voice, which contain fluctuations. Although train 10 was instrumental and did not contain fluctuation of pitch, the accuracy ratio was quite high, because the melody notes were not sustained for a long time, it seems. Our method failed on other clips because the melodies did not suffice our assumptions.

5. CONCLUSION

In this paper, we described a method to enhance melodic components in music audio signals, and a subsequent melody tracking algorithm. The enhancement method detected temporal-variability – fluctuation and shortness – of melodic components on differently resoluted two power spectrogram domain. The method comprised two stage HPSS-s, one was separation between "chordal" and "melodic + percussive," the other was between "chordal + melodic" and "percussive." Although the subsequent tracking was simple, the performance was quite high, thanks to the enhancement stage.

Our future works include constructing a melody-absence model and a duet (or more) model which we excluded in this paper, by using other available features of melody, e.g., harmonic structure of sound, timbrel information, and musicological knowledge, etc. Further, positive use of the other half component obtained in two-stage HPSS, "melody-suppressed signal," may also improve the performance. That will also be our future work.

6. REFERENCES

[1] M. Goto, "A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world signals," *Speech Communication*, vol. 43, pp. 311–329, 2004.

[2] C. Cao, M. Li, J. Liu, and Y. Yan, "Singing melody extraction in polyphonic music by harmonic tracking," *Proceedings of ISMIR*, 2007.

[3] J. S. Downie, "The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research," *Acoustical Science and Technology*, vol. 29, no. 4.

[4] J. L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," *Proceedings of ICASSP*, pp. 169–172, 2008.

[5] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," *Proceedings of ISMIR*, pp. 139–144, 2008.

[6] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," *Proceedings of EUSIPCO*, 2008.

[7] J. C. Brown, "An efficient algorithm for the calculation of a constant Q transform," *Journal of Acoustic Society of America*, vol. 92, no. 5, pp. 2698–2701, 1992.

[8] "http://labrosa.ee.columbia.edu/projects/melody/," .