Singing Voice Enhancement for Monaural Music Signals Based on Multiple Time-Frequency Analysis

Hideyuki Tachibana[1]*, Nobutaka, Ono*[1]*, Shigeki Sagayama*[1]

[1]Graduate School of Information Science and Technology,
The University of Tokyo, Japan.
{tachibana,onono,sagayama}@hil.t.u-tokyo.ac.jp

Abstract

We propose a novel technique to enhance singing voice in monaural music audio signals by capturing fluctuation of singing voice on spectrogram. Based on multiple spectrogram representation, the method separates an input signal into three components: stationary, fluctuated, and transient components, and singing voice is mainly included in the fluctuated component. The proposed algorithm consists of two stage processing of the sinusoidal/non-sinusoidal separation algorithm which we have recently developed. It is called harmonic/percussive sound separation (HPSS). In first stage, we filter out the stationary component based on HPSS analysis with long frame, and in second stage, we filter out the transient component based on HPSS analysis with short frame. We show that the proposed method effectively enhances the singing voice in music by experiments and show its application to melody extraction, which also supports the effectiveness of the method.

Index Terms: singing voice enhancement, singing voice suppression, harmonic/percussive sound separation, fluctuation, time-frequency resolution

1. Introduction

Music, especially popular music, often consists of melody sung by a singer and musical accompaniments played by instruments. In those music, vocal tracks are often the most impressive in all sounds. Therefore, in many music applications, e.g., automatic karaoke generation, singer identification, melody estimation, lyrics transcription, and in many music information retrieval (MIR)-related tasks, a technique to separate a music signal into vocal components and non-vocal components has much importance.

One of difficulties in singing voice enhancement is derived from the properties of "noise" to be reduced. In singing voice enhancement, "noise" consists of musical sounds played by musical instruments, e.g., guitar, drums, which we cannot assume several properties, which is assumed in other fields, such as whiteness, stationarity, noncorrelatedness to the signal, etc.

There have been several studies on singing voice extraction in music. Ozerov *et al.* [1] modeled spectra of singing voice by Gaussian mixture model, and derived an singing voice extraction method using a filter adaptation method. Some state-of-the-art singing voice enhancement algorithm consist of several stages. Li and Wang [2] proposed a singing voice enhancement method that is comprised of three stages: singing voice detection, predominant pitch detection, and filtering. Hsu and Jang's method [3] is also a staged processing. The method first discriminate accompaniment, unvoiced, and voiced segments, then track the predominant pitch using Li and Wang's method, and finally resynthesize the singing voice component.

Singing melody pitch extraction is also deeply related to melody extraction because pitch information is very useful for extracting melody. Goto [4] proposed a method called PreFEst which estimates the predominant pitch based on multi-agent model. Such F_0 estimation techniques can also be applied to enhance singing voice. In fact, for example, Fujihara *et al* [5] extracted harmonic structure of the pitch contour estimated by PreFEst, and used it for a MIR application.

Meanwhile, our approach to the problem is quite different from those of state-of-the-art methods. We focus on fluctuation of singing voice such as vibrato. In this paper, we show that the difference among stationary, fluctuated, and transient signals on spectrogram with various frame length. Then, we describe an algorithm to separate those three components based on a sinusoidal/non-sinusoidal separation algorithm, called harmonic/percussive sound separation (HPSS) [6], and we show some experimental results by using real-world music.

2. Spectrogram of Fluctuated Signal

2.1. Frame Length of STFT Analysis

In short-time Fourier transform (STFT) analysis, frame length is one of the significant parameters. Since it determines both of time and frequency resolution, resultant spectrogram representation would be very different as changing the frame length. In most signal analysis, one appropriate frame length suitable with signal of interest is chosen (e.g. typically 20-30ms in speech analysis). While, our aim here is to capture the fluctuated nature of singing voice by exploiting multiple spectrograms obtained from STFT with different frame lengths.

To start with, we consider a case of simple three types of signals, (1) very stationary sinusoidal wave, (2) a little fluctuating wave, and (3) non-stationary transient wave (Figure 1 Top). All three signals have many STFT representations, typically two STFT representations: STFT with short frame length, and with long frame length.

Let us think of a case of short frame length. In that case, the stationary sinusoidal signal can be represented similarly in neighboring time frames, i.e., it is continuous in time axis direction, while it is discontinuous in frequency axis direction. The fluctuating signal can also be represented similarly to the stationary sinusoidal signal, because the length of frame is short enough to ignore the fluctuation of the signal. On the other hand, the transient signal cannot be continuous in time axis direction, while it is continuous in frequency axis direction because of its impulsive nature. (Figure 1 Middle). Therefore, if we can define a criterion about those features of sounds, we can discriminate stationary sinusoidal signal and fluctuated signal

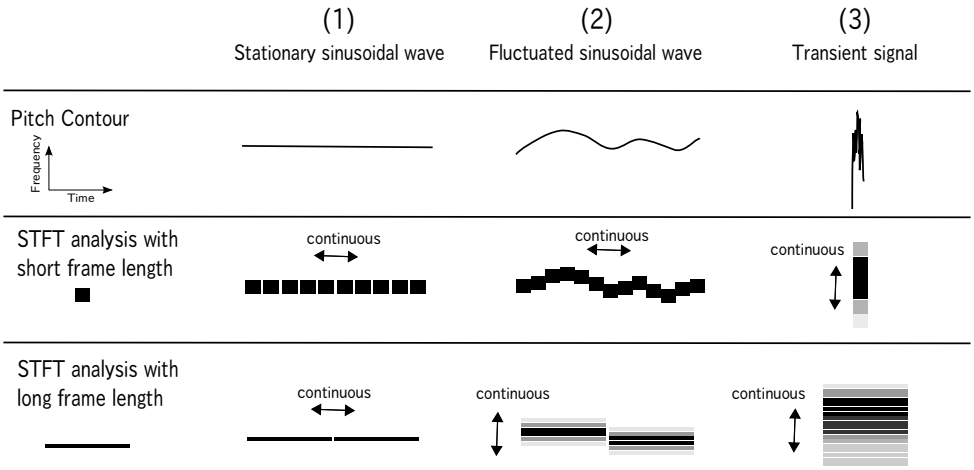

Figure 1: A concept of STFT representation of three types of signals. (1) Stationary sinusoidal signal can be continuous in time axis direction, discontinuous in frequency axis direction, on STFT spectrogram domain with both short and long frames (left middle and left bottom). (2) Fluctuated signal can be continuous in time axis direction, discontinuous in frequency axis direction on STFT spectrogram domain with short frame (center middle), while it is discontinuous in time axis direction and continuous in frequency axis direction on STFT spectrogram domain with long frame (center bottom). (3) Transient signal is almost always discontinuous in time axis direction, and continuous in frequency axis direction, regardless of the length of the frame (right middle and right bottom).

from transient signal.

Next, let us think of a case of long frame length. In that case, we can describe the stationary sinusoidal signal and transient signal similarly to the case of short frame length. However, the fluctuated signal needs to be described in another way. The fluctuating signal, in this case, cannot be represented similarly to the case of short frame length, because the length of frame is not short enough to ignore the fluctuation, but the effect of non-stationarity of the signal emerges within a frame of the spectrogram as broadness of the bandwidth. Besides, because the signal is fluctuating, a frame of the spectrogram is less likely to be similar to those of the neighbouring frames. Consequently, the spectrogram of the fluctuated signal is likely to be less continuous in time axis direction, while it is continuous in frequency direction because of the broadness of the bandwidth (Figure 1 Bottom). Therefore, the property of fluctuated signal is similar to that of transient signal in this case, and we can discriminate stationary sinusoidal signal from fluctuated signal and transient signal in a same manner.

2.2. STFT Representation of Singing Voice Signal

Here, we consider a case of music. Assume that we have a singing voice signal, and let us cut out a 250[ms] segment from the signal. In this case, because the segment is long enough, the non-stationarity of singing voice can be covered in a frame, and the shape of the spectrogram should be wide-band, and the spectrogram should be discontinuous in time axis direction. Similarly, let us think about the case of a segment length of which is 30[ms]. In this case, because the signal can be stationary enough in the segment, the shape of the spectrogram should be narrow-band, and continuous in time axis direction.

3. Multi-stage HPSS

3.1. Harmonic/Percussive Sound Separation

The characteristics of those spectral shapes of singing voice can be detected by an algorithm called harmonic and percussive sound separation (HPSS)[6]. In this section, we make a brief introduction of HPSS.

The original purpose of HPSS was to separate a music signal into "harmonic" components $\mathbf{H} = \{H_{t,\omega}\}_{0\leq t\leq T, 0\leq k\leq K}$ and "percussive" components $\mathbf{P} = \{P_{t,\omega}\}_{0\leq t\leq T, 0\leq k\leq K}$, where t and k are time and frequency indices. More precisely, the algorithm separates a STFT spectrogram $\mathbf{W} = \{W_{t,\omega}\}_{0\leq t\leq T, 0\leq k\leq K}$ into a spectrogram $\mathbf{H}$ and $\mathbf{P}$, which are smooth in time axis direction and frequency axis direction respectively, under a constraint that sum of $\mathbf{H}$ and $\mathbf{P}$ should be almost equal to the original spectrogram $\mathbf{W}$,

HPSS is formulated as an optimization problem. The objective function to be minimized is

$$\begin{aligned} J[\mathbf{H},\mathbf{P}] &= \frac{1}{\sigma_H^2}\sum_{t=0}^{T-1}\sum_{k=0}^{K}(\sqrt{H_{t+1,k}}-\sqrt{H_{t,k}})^2 \\ &+ \frac{1}{\sigma_P^2}\sum_{t=0}^{T}\sum_{k=0}^{K-1}(\sqrt{P_{t,k+1}}-\sqrt{P_{t,k}})^2 \\ &+ \mathcal{I}(\mathbf{W},\mathbf{H}+\mathbf{P}), \end{aligned} \tag{1}$$

where $\mathcal{I}(\cdot)$ denotes $\mathcal{I}$-divergence. The first term of the equation is the log-likelihood function of $\mathbf{H}$: the spectrogram of H is likely to be smooth in time axis direction. The second term of the equation is that of $\mathbf{P}$ similarly: the spectrogram of P is likely to be smooth in frequency axis direction. The third term of the equation constrains sum of $\mathbf{H}$ and $\mathbf{P}$ to be close to the original signal $\mathbf{W}$.

The objective function can be optimized by EM-like algorithm. Thus we can separate a spectrogram $\mathbf{W}$ into $\mathbf{H}$ and $\mathbf{P}$, i.e.,

$$\mathbf{W} \xrightarrow{\text{HPSS}} \{\mathbf{H},\mathbf{P}\}, \tag{2}$$

and by applying inverse STFT to the obtained spectrogram, we can obtain audible signals $h(t)$ and $p(t)$, as shown in Figure 2. Note that the length of frame of STFT is arbitrary, and obtained $h(t)$ and $p(t)$ are not unique, i.e., they also have frame length l_k as a parameter; $h_{l_k}(t), p_{l_k}(t)$.

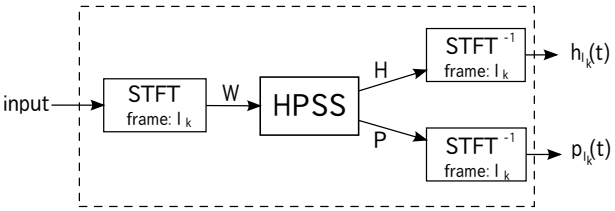

Figure 2: Diagram of a block of HPSS processing.

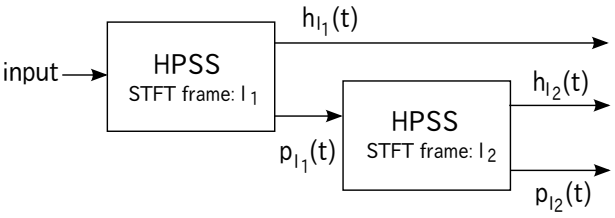

Figure 3: Diagram of multi-stage HPSS.

3.2. Multi-stage HPSS

Despite the original purpose of HPSS, the algorithm can separate not only "harmonic" and "percussive" components, but also separate e.g. "singing voice" and "harmonic sounds," because of the reason that we described in section 2.

Actually, by adjusting the frame length of STFT, we can regulate into which a signal is separated, H or P. When we run HPSS on long-frame STFT domain, singing voice can be separated into P, while if on short-frame STFT domain, singing voice can be separated into H, because of the natures of singing voice mentioned above. Therefore, by applying HPSS like a filter, we can obtain singing-voice-enhanced signals. We call the two-stage processing as "multi-stage HPSS."

The procedure of multi-stage HPSS is as follows: we first transform the signal by STFT using a long frame l_1, and separate a signal on the STFT domain,

$$\mathbf{W}_{l_1} \xrightarrow{\text{HPSS}} \{\mathbf{H}_{l_1}, \mathbf{P}_{l_1}\}, \tag{3}$$

then reconstruct the waveform $p_1(t)$ from $\mathbf{P}_{l_1}$, and calculate its STFT using short frames, length of which is l_2,

$$\mathbf{P}_{l_1} \xrightarrow{\text{STFT}_{l_1}^{-1}} p_{l_1}(t) \xrightarrow{\text{STFT}_{l_2}} \mathbf{P}_{l_2}. \tag{4}$$

Finally, we separate the signal by HPSS,

$$\mathbf{P}_{l_2} \xrightarrow{\text{HPSS}} \{\mathbf{H}_{l_2}, \mathbf{P}_{l_2}\}. \tag{5}$$

Thus obtained $\text{STFT}_{l_2}^{-1}[\mathbf{H}_{l_2}] =: h_{l_2}(t)$ is the desired singing-voice-enhanced signal (Figure 3). The residual $\text{STFT}_{l_1}^{-1}[\mathbf{H}_{l_1}] =: h_{l_1}(t)$ and $\text{STFT}_{l_2}^{-1}[\mathbf{P}_{l_2}] =: p_{l_2}(t)$ can be also used for some applications, though we do not use them in this paper.

4. Experiments

4.1. Singing Voice Enhancement

We applied multi-stage HPSS to real-world musics, e.g., LabROSA dataset [7]. The data were 16000kHz monaural PCM data. The parameters we used were as follows: analysing and reconstructing window of STFT was sine window (square root of hanning window), with half overlapping. This arrangement satisfies the condition that $(\text{STFT}^{-1} \cdot \text{STFT})$ is identical. The length of frame were 32 [ms] (512 samples) and 256 [ms] (4096 samples).

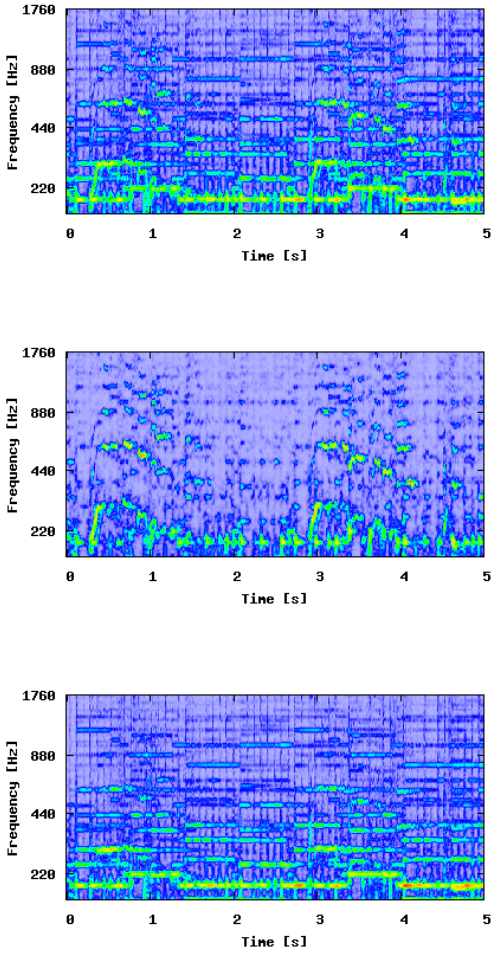

Figure 4: Spectrograms of input signal, excerpted from LabROSA dataset [7] (top), singing-voice-enhanced signal $h_{l_2}(t)$ (middle), and singing-voice-suppressed signal $h_{l_1}(t) + p_{l_2}(t)$ (bottom).

Figure 4 shows a result of the singing voice enhancement. Comparing the top spectrogram and the middle spectrogram, we can see that the singing voice component in middle spectrogram appears clearer than that of top spectrogram, i.e., the method suppress the accompanying signals effectively. The bottom spectrogram is the sum of the residual signals $h_{l_1}(t) + p_{l_2}(t)$ in Figure 3, in which there are little singing voice component, that shows that the method has a property that the method does not suppress singing voice component so much.

From our qualitative preliminary experiments, the method seemed to be effective especially in techno, rock, and pop music auditorily, and not effective enough in classical music, jazz, and enka. Besides, sounds of violin and trumpet tended to be recognized as "singing voice," and reverberation of singing voice tended to be recognized as "non-vocal."

4.2. MIR-related Application: Audio Melody Extraction

As mentioned in the introduction, singing voice enhancement is related to singing melody pitch extraction technique, and they can be used as another's preprocessing mutually. In this section, we show a result of a singing melody pitch extraction method, which uses the singing enhancement algorithm as a preprocessing.

The method is comprised of following two stages [8],

1. Apply the singing voice enhancement method.
2. Estimate the pitch contour by a simple tracking algorithm based on dynamic programming.

The method was evaluated in the framework of the Audio Melody Extraction (AME) evaluation in MIREX (Music Information Retrieval EXchange)[9]. Figure 5 shows the excerpted results from AME evaluation in MIREX 2009 [10] and MIREX 2010 [11], which show that the performance of our method (TOOS and TOOS1) is high, especially in a condition that the volume level of melody is relatively low (-5 dB) to accompanying instruments. The figure also shows that the performance of our method is also comparable to other methods in a condition that the volume level of melody is relatively high (+5 dB). The results show the effectiveness of our singing voice enhancement method as a preprocessing for singing melody pitch extraction.

5. Conclusions

In this paper, we described a novel method to enhance singing voice based on the fluctuation of the signal. The separation algorithm focused on the discriminative spectral shapes of fluctuated sound: it is continuous in time axis direction and discontinuous in frequency axis direction when analysed by STFT with short frame length, while it is discontinuous in time axis direction and continuous in frequency axis direction when analysed by STFT with long frame length. We showed an example of the result of the method using real-world music signals, and we also showed the effectiveness of the method as a preprocessing for an application: Audio Melody Extraction.

Our future works include investigation of an application of the method related to music information retrieval tasks, combination with some other music signal processing techniques, and investigation of an application, based on the residual accompanying signals, e.g., automatic karaoke generator, etc.

6. Acknowledgements

A part of this research was supported by Grant-in-Aid for JSPS Fellows (No. 22-6961).

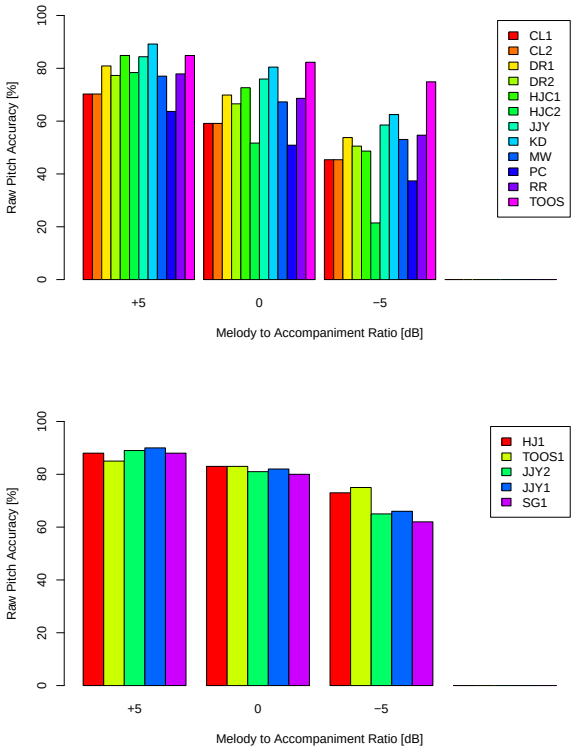

Figure 5: Excerpted results from AME evaluation, MIREX 2009 [10] (top) and MIREX 2010 [9] (bottom): Raw Pitch Accuracy (RPA) in +5dB, 0dB, -5dB SNR (i.e., melody to accompaniment ratio) conditions. TOOS and TOOS1 is our method. This graph shows that the performance of our method is high, especially in low dB conditions, and comparable to other methods in high dB conditions. It shows the effectiveness of our singing voice enhancement method as a preprocessing for singing melody pitch extraction.

7. References

[1] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One Microphone Singing Voice Separation Using Source-adapted Models," *Proc. WASPAA,* pp.90–93, 2005.

[2] Y. Li, and D. L. Wang, "Separation of Singing Voice From Music Accompaniment for Monaural Recordings," *IEEE Trans. ASLP,* Vol. 15, No. 4, pp. 1475–1487, 2007.

[3] C.-L. Hsu and J.-S. R. Jang, "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset," *IEEE Trans. ASLP,* Vol. 18, No. 2, pp. 310–319, 2010,

[4] 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, pp. 311–329, 2004.

[5] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Singer Identification Based on Accompaniment Sound Reduction and Reliable Frame Selection," *Proc. ISMIR,* pp.329–336 2005.

[6] 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.

[7] `http://labrosa.ee.columbia.edu/projects/melody/`

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

[9] 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, pp. 247–255, 2008

[10] `http://www.music-ir.org/mirex/wiki/2009:Audio_Melody_Extraction_Results`

[11] `http://www.music-ir.org/mirex/wiki/2010:MIREX2010_Results`