SINGING VOICE SEPARATION BASED ON SPARSE NATURE AND SPECTRAL/TEMPORAL DISCONTINUITY

Il-Young Jeong **Kyogu Lee**

Music and Audio Research Group, Seoul National University
{finejuly, kglee}@snu.ac.kr

ABSTRACT

This extended abstract describe our singing voice separation algorithm submitted to the Music Information Retrieval Evaluation eXchange (MIREX) 2014. This algorithm is based on the assumption all the non-vocal sources in a music signal can be classified into the harmonic or percussive instrument, while vocal has its unique characteristics.

1. INTRODUCTION

We propose a novel singing voice separation algorithm using its spectral/temporal discontinuity. This submission it basically an implementation of our recent paper [1].

2. ALGORITHM

Let us say we have a spectrogram of the music signal. We assume that it can be represented as a sum of the harmonic, percussive, and vocal components as

$$W = |\boldsymbol{X}|^{2\gamma} = H + P + V, \quad (1)$$

where $\boldsymbol{X}$ is a short-time Fourier transform of the music, $|\cdot|^{2\gamma}$ denotes the element-wise power operation, and the scale parameter γ, in the interval of (0, 1], denotes the compression rate. W, H, P, and V denote the scale compressed version of the magnitude spectrogram of the music, harmonic instrument, percussive instrument, and vocal, respectively.

On the spectrogram domain, harmonic and percussive sources show the distinguished characteristics. In case of harmonic sources, which has a strong harmonic structure and a long sustain time, it is shown as horizontal ridges, while percussive sources shown as vertical ridges because of its broadband spectra and short sustain time. Meanwhile, vocal has very unique characteristics. While it also has strong harmonic structure as harmonic instruments, this spectra is fluctuated very fast and it can be considered to have a short sustain time as percussive instruments.

Based on this observation, we derive the following objective function.

$$J(H,P,V) = \frac{1}{2}\sum_{f,t}\left(H_{f,t-1} - H_{f,t}\right)^2 + \frac{\alpha}{2}\sum_{f,t}\left(P_{f-1,t} - P_{f,t}\right)^2 + \varphi\sum_{f,t}\left|V_{f,t}\right|,$$
$$s.t. \quad H + P + V = W, \quad H, P, V \geq 0, \quad (2)$$

where f and t denote the frequency and the time indices, respectively. α denotes the relative weight between the minimization of the temporal/spectral gradient of the harmonic and percussive components, and φ is a relative weight of the vocal minimization. In general, higher value of φ increase the signal-to-interference ratio, while lower value increase the signal-to-artifact ratio. For the details of the optimization procedure, please refer [1].

To maximize the separation performance, a high-pass filter can be applied as a post-processing step. Since vocal is rarely distributed in the low-frequency, we remove all the separated singing voice lower than 100Hz, and added it to the separated accompaniment.

3. PARAMETERS

In this submission, we set the parameters as table 1. All the values are obtained empirically to maximize the signal-to-distortion ratio of the separated singing voice and the accompaniment both.

Table 1. Parameter setting

Parameter	γ	α	φ	Iteration
Value	0.25	0.25	0.01	200

4. REFERENCES

[1] I.-Y. Jeong and K. Lee, "Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints," *IEEE Signal Processing Letters*, Vol. 21, No. 10, pp. 1197-1200, 2014.