MIREX2014: SINGING VOICE SEPARATION

Yukara Ikemiya  Kazuyoshi Yoshii  Katsutoshi Itoyama
Department of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
{ikemiya, yoshii, itoyama}@kuis.kyoto-u.ac.jp

ABSTRACT
This paper describes our submission for the singing voice separation task of the Music Information Retrieval Evaluation eXchange (MIREX 2014).

1. INTRODUCTION
A typical approach to singing-voice separation is to estimate the vocal F0 contour from a target music signal and then extract the singing voice by using a time-frequency mask that passes only the harmonic components of the vocal F0s and overtones. Vocal F0 estimation, on the contrary, is considered to become easier if only the singing voice can be extracted accurately from the target signal. Such mutual dependency has scarcely been focused on in most conventional studies. To overcome this limitation, our framework alternates those two tasks while using the results of each in the other (Fig. 1). More specifically, we first extract the singing voice by using robust principal component analysis (RPCA) [1]. The F0 contour is then estimated from the separated singing voice by finding the optimal path over a F0-saliency spectrogram based on sub-harmonic summation (SHS). This enables us to improve singing-voice separation by combining a time-frequency mask based on RPCA with a mask based on harmonic structures.

2. METHOD
2.1 First-stage singing voice separation
One of the most promising methods for singing voice separation is to focus on the repeating nature of accompanying sounds [1, 2]. The difference between vocal and accompanying sounds is well characterized in the time-frequency domain. Since the timbres of harmonic instruments, such as pianos and guitars, are consistent for each pitch and the pitches are basically discretized at a semitone level, harmonic spectra having the same shape appear repeatedly in the same musical piece. The spectra of unpitched instruments (e.g., drums) also tend to appear repeatedly. Vocal spectra, in contrast, rarely have the same shape because the timbres and pitches of vocal sounds vary significantly and continuously over time.

In this submission, we use robust principal component analysis (RPCA) to separate non-repeating components, as vocal sounds, from a polyphonic spectrogram [1]. When RPCA is applied to the STFT\(^1\) spectrogram of a polyphonic music signal, spectral components having repeating structures are allocated to a low-rank matrix and the other varying components are allocated to a sparse matrix. Then a time-frequency binary mask is made by comparing each element of \(L\) with the corresponding element of \(S\). The sung melody is extracted by applying the binary mask to the original spectrogram.

2.2 Melody extraction
2.2.1 Salience function
SHS [3] is a simple algorithm that underlies many melody extraction methods. A salience function \(H(t, s)\) is formulated on a logarithmic scale as follows:

\[
H(t, s) = \sum_{n=1}^{N} h_n P(t, s + 1200 \log_2 n),
\]

where \(t\) and \(s\) indicate a frame index and a logarithmic frequency [cents], respectively, \(P(t, s)\) represents the power at frame \(t\) and frequency \(s\), \(N\) is the number of harmonic partials considered, and \(h_n\) is a decaying factor (0.86\(^n\) in this submission). The log-frequency power spectrum \(P(t, s)\) is calculated from the STFT spectrum via spline interpolation. The frequency resolution of \(P(t, s)\) is 200 bins per octave (6 cents per bin). Before computing the salience function, we apply to the original spectrum the A-weighting function\(^2\), which takes into account the non-linearity of human auditory perception.

\(^1\) short-time Fourier transform
\(^2\) http://replaygain.hydrogenaud.io/proposalequal_loudness.html
2.2.2 Viterbi search
Given a salience function as a time-frequency spectrogram, we estimate the optimal melody contour \( \hat{S} \) by solving an optimal path problem formulated as follows:

\[
\hat{S} = \arg\max_{s_1, \ldots, s_T} \sum_{t=1}^{T-1} \left\{ \log a_t H(t, s_t) + \log T(s_t, s_{t+1}) \right\},
\]

(2)

where \( T(s_t, s_{t+1}) \) is a transition probability that indicates how likely the current F0 \( s_t \) is to move on to the next F0 \( s_{t+1} \), and \( a_t \) is a normalization factor that makes the salience values sum to 1 within a range of F0 search. \( T(s_t, s_{t+1}) \) is given by the Laplace distribution, \( \mathcal{L}(s_t - s_{t+1} | 0, 150) \), with a zero mean and a standard deviation of 150 cents. The time frame interval is 10 msec. Optimal \( \hat{S} \) can be effectively found by using the Viterbi search.

2.3 Singing voice separation based on vocal F0s
Assuming that vocal spectra preserve their original harmonic structures and the energy of those spectra is localized on harmonic partials after singing voice separation based on RPCA, we make a binary mask \( M_h \) that passes only harmonic partials of given vocal F0s:

\[
M_h(t, f) = \begin{cases} 
1 & \text{if } nF_t - \frac{w}{2} < f < nF_t + \frac{w}{2}, \\
0 & \text{otherwise},
\end{cases}
\]

(3)

where \( F_t \) is the vocal F0 estimated from frame \( t \), \( n \) is the index of a harmonic partial, and \( w \) is a frequency width for extracting the energy around each harmonic partial.

We integrate the harmonic mask \( M_h \) with the binary mask \( M_r \) obtained using the RPCA-based method. Finally, the vocal spectra \( P_v \) and the accompanying spectra \( P_a \) are given by

\[
P_v(t, f) = M_h(t, f) M_r(t, f) P(t, f),
\]

\[
P_a(t, f) = P(t, f) - P_v(t, f),
\]

(4)

where \( P \) is the original spectrogram of a polyphonic music signal. The separated vocal signals and accompanying signals are obtained by calculating the inverse STFT of each of the spectra.

2.4 Vocal activity detection
We apply simple vocal activity detection (VAD) based on thresholding. First we design a cost function for thresholding as follows.

\[
\text{CF}(t) = \sum_f \left\{ \frac{1}{H_f} \sum_{s=1}^{H_f} P(t, s + 1200 \log_2 n) \right\}^{1.8}
\]

(5)

where \( H_f \) is the number of all harmonics within 4000 [Hz] for each frequency \( f \). Using this function, vocal and non-vocal state are estimated by thresholding.

\[
s_t = \begin{cases} 
s_v & \text{if } \text{CF}(t) > k \\
s_n & \text{otherwise}
\end{cases}
\]

(6)

where \( k \) is a threshold.