REPET-SIM FOR SINGING VOICE SEPARATION

Zafar Rafii
Northwestern University
Department of EECS

Bryan Pardo
Northwestern University
Department of EECS

ABSTRACT

This extended abstract presents the REPET-SIM algorithm for the task of singing voice separation for the Music Information Retrieval Evaluation eXchange (MIREX) 2014. REPET-SIM is a generalization of the REpeating Pattern Extraction Technique (REPET) that uses a similarity matrix to identify and separate the repeating background from the non-repeating foreground in an audio mixture. The method assumes that, in audio, mixtures can often be understood as a background component that is generally more repeating in time, superimposed with a foreground component that is generally more variable in time (e.g., a song with varying vocals overlaid on a repeating accompaniment, or a recording with a varying speech mixed up with a repeating noise). The basic idea is to identify repeating elements in the mixture by measuring self-similarity along time, derive repeating models by averaging the repeating elements over their repetitions, and extract the repeating structure by comparing the repeating models to the mixture. For more details, including source codes, audio examples, and referred publications, please see `http://music.eecs.northwestern.edu/research.php?project=repet`.

1. METHOD

REPET-SIM can be summarized in three stages (see Figure 1): (1) identification of the repeating indices (see Section 1.1), (2) modeling of a repeating spectrogram (see Section 1.2), and (3) extraction of the repeating structure (see Section 1.3) [3, 4].

1.1 Repeating Elements Identification

Similarities in a signal can be found by using the similarity matrix, which is a two-dimensional representation where each point (a, b) measures the (dis)similarity between any two elements a and b of a given sequence.

Given a mixture signal x, we first compute its Short-Time Fourier Transform (STFT) X using windows of N samples. We then derive the magnitude spectrogram V by taking the absolute value of the elements of X, after discarding the complex conjugates (i.e., the frequency channels above half the sampling frequency).

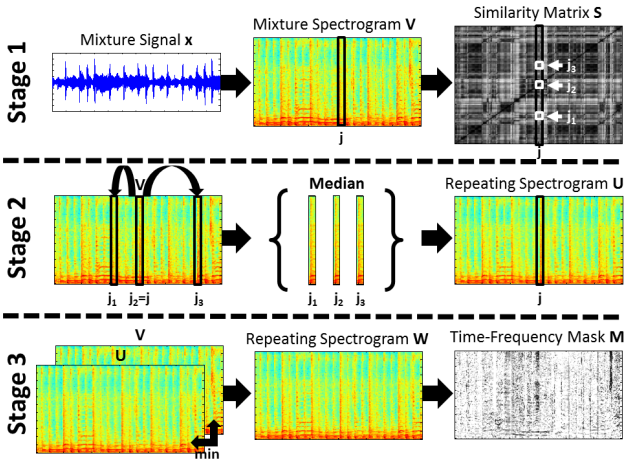

Figure 1. Overview of REPET-SIM. **Stage 1**: computation of the similarity matrix S and identification of the repeating indices j_k's. **Stage 2**: filtering of the mixture spectrogram V and computation of the initial repeating spectrogram U. **Stage 3**: computation of the refined repeating spectrogram W and derivation of the time-frequency mask M.

We then derive the similarity matrix S by computing the cosine similarity between all the pairs of time frames j_a and j_b in V. We chose the cosine similarity because it is a similarity measure that is commonly used and conveniently normalized (i.e., $\in [0, 1]$). The computation of the similarity matrix S is shown in Equation 1.

$$S(j_a, j_b) = \frac{\sum_{i=1}^{n} V(i, j_a) V(i, j_b)}{\sqrt{\sum_{i=1}^{n} V(i, j_a)^2} \sqrt{\sum_{i=1}^{n} V(i, j_b)^2}} \quad (1)$$

for $i = 1 \dots n$ and $j_a, j_b = 1 \dots m$

where $n = \frac{N}{2} + 1$ = number of frequency channels

and m = number of time frames

The idea of the similarity matrix was introduced in [1], except that the magnitude spectrogram and the cosine similarity are used here in lieu of the Mel Frequency Cepstrum Coefficients (MFCC) and the dot product, respectively. Pilot experiments showed that this method allows for a clearer visualization of the underlying repeating structure in the mixture.

Once the similarity matrix S is computed, we use it to identify the repeating indices. If repeating elements are present in the mixture, S would form regions of high and low similarity at different indices, unveiling the underlying repeating structure of the mixture, as exemplified in the top row of Figure 1. We then identify, for every time frame j in the mixture spectrogram V, the time frames j_k's that are the most similar to j using S, and save their indices in a vector J_j. We use the following constraint parameters: t, the minimum similarity between a time frame j_k and time frame j; d, the minimum distance between two consecutive time frames j_k; and k, the maximum number of time frames j_k.

The computation of the similarity matrix S and the identification of the repeating indices j_k's are illustrated in the top row of Figure 1.

1.2 Repeating Spectrogram Modeling

Once the repeating indices j_k's are identified, we use them to derive an initial repeating spectrogram U. For every time frame j in the mixture spectrogram V, we derive the corresponding time frame j in U by taking, for every frequency channel, the median of the k time frames repeating at indices j_k given by the vector J_j, where k is the maximum number of repeating time frames, as exemplified in the middle row of Figure 1. The computation of the initial repeating spectrogram U is shown in Equation 2.

$$\begin{aligned} &U(i,j) = \underset{l=1\ldots k}{median}\{V(i, J_j(l)\} \\ &\text{for } i = 1\ldots n \quad \text{and } j = 1\ldots m \\ &\text{where } J_j = \{j_1 \ldots j_k\} = \text{repeating indices for time frame } j \\ &\text{and } k = \text{maximum number of repeating time frames} \end{aligned} \tag{2}$$

The rationale is that, if we assume that the non-repeating foreground has a sparse and varied time-frequency representation compared with the time-frequency representation of the repeating background, time-frequency bins with small deviations between their indices j_k's in V would most likely represent repeating elements, which would be captured by the median filter. On the other hand, time-frequency bins with large deviations between their indices j_k's in V would most likely be corrupted by non-repeating elements (i.e., outliers), which would be removed by the median filter.

The filtering of the mixture spectrogram V and the computation of the initial repeating spectrogram U are illustrated in the middle row of Figure 1.

Compared with the REPET methods (the original and the extensions) that depend on periodicity [2, 5], REPET-SIM depends on similarity, so that it can handle non-periodically repeating structures, i.e., when the repeating patterns happen intermittently or without a clear periodicity.

1.3 Repeating Structure Extraction

Once the initial repeating spectrogram U is computed, we use it to derive a refined repeating spectrogram W, by taking the element-wise minimum between U and the mixture spectrogram V, as exemplified in the bottom row of Figure 1. The computation of the refined repeating spectrogram W is shown in Equation 3.

$$\begin{aligned} &W(i,j) = \min\left\{U(i,j), V(i,j)\right\} \\ &\text{for } i = 1\ldots n \quad \text{and} \quad j = 1\ldots m \end{aligned} \tag{3}$$

Once the refined repeating spectrogram W is computed, we use it to derive a time-frequency mask M, by dividing W by the mixture spectrogram V, element-wise. A time-frequency mask is designed to filter the energy of a spectrogram by multiplying each time-frequency bin by a weight value (typically $\in$ [0,1]); it is commonly used for source separation. The computation of the time-frequency mask M is shown in Equation 4.

$$\begin{aligned} &M(i,j) = \frac{W(i,j)}{V(i,j)} \\ &\text{for } i = 1\ldots n \quad \text{and} \quad j = 1\ldots m \end{aligned} \tag{4}$$

The computation of the refined repeating spectrogram W and the derivation of the time-frequency mask M are illustrated in the bottom row of Figure 1.

The time complexity of REPET-SIM is $O(m^2)$, where m is the number of time frames in the spectrogram. The computation of the similarity matrix takes $O(m^2)$, as it is based on a matrix multiplication, while the median filtering takes $O(m)$. The time complexity of the original REPET [5] and the adaptive REPET [2] are $O(m \log m)$.

2. REFERENCES

[1] Jonathan Foote. Visualizing music and audio using self-similarity. In *7th ACM International Conference on Multimedia*, pages 77–80, Orlando, FL, USA, October 30-November 5 1999.

[2] Antoine Liutkus, Zafar Rafii, Roland Badeau, Bryan Pardo, and Gaël Richard. Adaptive filtering for music/voice separation exploiting the repeating musical structure. In *37th International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 25-30 2012.

[3] Zafar Rafii, Antoine Liutkus, and Bryan Pardo. REPET for background/foreground separation in audio. In Ganesh R. Naik and Wenwu Wang, editors, *Blind Source Separation*, Signals and Communication Technology, chapter 14, pages 395–411. Springer Berlin Heidelberg, 2014.

[4] Zafar Rafii and Bryan Pardo. Music/voice separation using the similarity matrix. In *13th International Society for Music Information Retrieval*, Porto, Portugal, October 8-12 2012.

[5] Zafar Rafii and Bryan Pardo. REpeating Pattern Extraction Technique (REPET): A simple method for music/voice separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(1):71–82, January 2013.