Chapter 14
REPET for Background/Foreground Separation in Audio

Zafar Rafii, Antoine Liutkus and Bryan Pardo

Abstract Repetition is a fundamental element in generating and perceiving structure. In audio, mixtures are often composed of structures where a repeating background signal is superimposed with a varying foreground signal (e.g., a singer overlaying varying vocals on a repeating accompaniment or a varying speech signal mixed up with a repeating background noise). On this basis, we present the *REpeating Pattern Extraction Technique (REPET)*, a simple approach for separating the repeating background from the non-repeating foreground in an audio mixture. The basic idea is to find the repeating elements in the mixture, derive the underlying repeating models, and extract the repeating background by comparing the models to the mixture. Unlike other separation approaches, REPET does not depend on special parameterizations, does not rely on complex frameworks, and does not require external information. Because it is only based on repetition, it has the advantage of being simple, fast, blind, and therefore completely and easily automatable.

14.1 Introduction

Figure–ground perception is the ability to segregate a scene into a foreground component (figure) and a background component (ground). In vision, the most famous example is probably the *Rubin vase*: depending on one's attention, one would perceive either a vase or two faces [19]. In auditory scene analysis [2], different cues

Z. Rafii (✉) · B. Pardo
Northwestern University, Evanston, IL, USA
e-mail: zafarrafii@u.northwestern.edu

B. Pardo
e-mail: pardo@northwestern.edu

A. Liutkus
Inria, PAROLE, Villiers-lès-Nancy, France
e-mail: antoine.liutkus@inria.fr

G. R. Naik and W. Wang (eds.), *Blind Source Separation*,
Signals and Communication Technology, DOI: 10.1007/978-3-642-55016-4_14,

can be used to segregate foreground and background: loudness (e.g., the foreground signal is louder), spatial location (e.g., the foreground signal is in the center of the stereo field), or timbre (e.g., the foreground signal is a woman speaking).

Unlike fixed images (e.g., Rubin vase), audio has also a temporal dimension that can be exploited for segregation. Particularly, auditory scenes are often composed of a background component that is more stable or repeating in time (e.g., air conditioner noise or footsteps), and a foreground component that is more variable in time (e.g., a human talking or a saxophone solo). The most notable examples are probably seen (or rather heard) in music. Indeed, musical works are often organized into structures where a varying melody is overlaid on a repeating background (e.g., rapping over a repeating drum loop or playing a solo over a repeating chord progression). This implies that there should be patterns repeating in time that could be used to discriminate the background from the foreground in an auditory scene.

Repetition also appears as an exploitable cue for source separation in audio. By identifying and extracting the repeating patterns (e.g., drum loop or guitar riff), we show that it is possible to separate the repeating background from the non-repeating foreground in an audio mixture. This idea is supported by recent findings in cognitive psychology which showed that human listeners are able to segregate individual audio sources if they repeat across different mixtures, even in the absence of other cues (e.g., spatial location) and without a prior knowledge of the sources [10].

In this chapter, we present the *REpeating Pattern Extraction Technique (REPET)*, a simple method that uses repetition as a basis for background/foreground separation in audio. The basic idea is to find the repeating elements in the mixture, derive the underlying repeating models, and extract the repeating background by comparing the models to the mixture. The rest of this chapter is organized as follows.

In Sect. 14.2, we present the original REPET. The original REPET aims at identifying and extracting the repeating patterns in an audio mixture, by estimating a period of the underlying repeating structure and modeling a segment of the periodically repeating background [13, 16]. The idea can be loosely related to background subtraction, a technique used in computer vision for separating moving foreground objects from a fixed background scene in a sequence of video frames [12].

In Sect. 14.3, we present the adaptive REPET. The original REPET works well when the repeating background is relatively stable (e.g., a verse or the chorus in a song); however, the repeating background can also vary over time (e.g., a verse followed by the chorus in the song). The adaptive REPET is an extension of the original REPET that can handle varying repeating structures, by estimating the time-varying repeating periods and extracting the repeating background locally, without the need for segmentation or windowing [9].

In Sect. 14.4, we present *REPET-SIM*. The REPET methods work well when the repeating background has periodically repeating patterns (e.g., jackhammer noise); however, the repeating patterns can also happen intermittently or without a global or local periodicity (e.g., frogs by a pond). REPET-SIM is a generalization of REPET that can also handle non-periodically repeating structures, by using a similarity matrix to identify the repeating elements [14, 15].

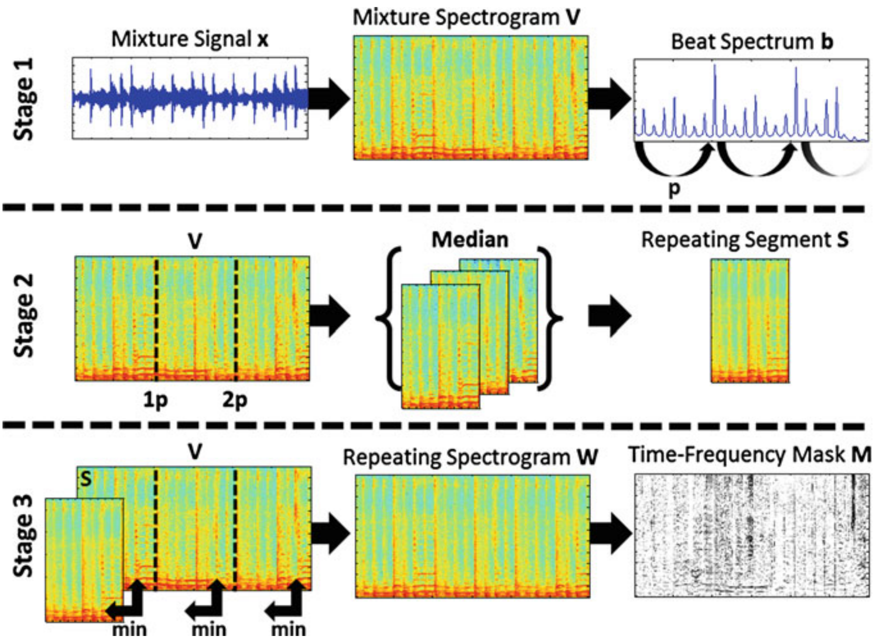

Fig. 14.1 Overview of the original REPET. *Stage 1* calculation of the beat spectrum b and estimation of a repeating period p. *Stage 2* segmentation of the mixture spectrogram V and calculation of the repeating segment model S. *Stage 3* calculation of the repeating spectrogram model W and derivation of the soft time–frequency mask M

14.2 REpeating Pattern Extraction Technique

The original REPET aims at identifying and extracting the repeating patterns in an audio mixture, by estimating a period of the underlying repeating structure and modeling a segment of the periodically repeating background [13, 16].

The original REPET can be summarized in three stages (see Fig. 14.1): (1) identification of a repeating period (see Sect. 14.2.1), (2) modeling of a repeating segment (see Sect. 14.2.2), and (3) extraction of the repeating structure (see Sect. 14.2.3).

14.2.1 Repeating Period Identification

Periodicities in a signal can be found by using the autocorrelation, which is the cross-correlation of a signal with itself. The function basically measures the similarity between a segment and a lagged version of itself over successive time lags.

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 symmetric part (i.e., the frequency channels above half the sampling frequency).

We then compute the autocorrelation over time for each frequency channel of the power spectrogram V^2 (i.e., the element-wise square of V) and obtain the matrix of autocorrelations A. We use V^2 to emphasize peaks of periodicity in A. If x is stereo, V^2 can be averaged over the channels. The overall self-similarity b of x is then obtained by taking the mean over the rows of A. We finally normalize b by dividing it by its first term (i.e., time lag 0). The calculation of b is shown in Eq. 14.1.

$$
\begin{aligned}
A(i,l) &= \frac{1}{m-l+1} \sum_{j=1}^{m-l+1} V(i,j)^2 V(i,j+l-1)^2 \\
b(l) &= \frac{1}{n} \sum_{i=1}^{n} A(i,l) \quad \text{then } b(l) = \frac{b(l)}{b(1)} \qquad (14.1) \\
&\text{for } i = 1 \ldots n \quad \text{where } n = \frac{N}{2} + 1 = \text{number of frequency channels} \\
&\text{for } l = 1 \ldots m \quad \text{where } m = \text{number of time frames.}
\end{aligned}
$$

The idea is very similar to the *beat spectrum* introduced in [7], except that no similarity matrix is explicitly calculated here, and the dot product is used in lieu of the cosine similarity. Pilot experiments showed that this method allows for a clearer visualization of the underlying periodically repeating structure in the mixture. For simplicity, we will refer to b as the beat spectrum for the remainder of this chapter.

Once the beat spectrum b is calculated, the first term which measures the similarity of the whole signal with itself (i.e., time lag 0) is discarded. If periodically repeating patterns are present in x, b would form peaks that are periodically repeating at different period rates, unveiling the underlying periodically repeating structure of the mixture, as exemplified in the top row of Fig. 14.1.

We then use a period finder to estimate the repeating period p from b. One approach can be to identify the period in the beat spectrum that has the highest mean accumulated energy over its integer multiples (see Algorithm 1 in [16]). Another approach can be to find the local maximum in a given lag range of the beat spectrum (see source codes online[1]).

The calculation of the beat spectrum b and the estimation of the repeating period p are illustrated in the top row of Fig. 14.1.

14.2.2 Repeating Segment Modeling

Once the repeating period p is estimated, we use it to segment the mixture spectrogram V into r segments of length p. We then take the element-wise median of the r

[1] http://music.eecs.northwestern.edu/research.php?project=repet

segments and obtain the repeating segment model S, as exemplified in the middle row of Fig. 14.1. The calculation of the repeating segment model S is shown in Eq. 14.2.

$$\begin{aligned}
&S(i, j) = \underset{k=1...r}{\text{median}}\left\{V(i, j + (k-1)p)\right\} \\
&\text{for } i = 1 \ldots n \quad \text{and} \quad j = 1 \ldots p \\
&\text{where } p = \text{period length} \quad \text{and} \quad r = \text{number of segments.}
\end{aligned} \tag{14.2}$$

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 at period rate p would most likely represent repeating elements and would be captured by the median model. On the other hand, time–frequency bins with large deviations at period rate p would most likely be corrupted by non-repeating elements (i.e., outliers) and would be removed by the median model.

The segmentation of the mixture spectrogram V and the calculation of the repeating segment model S are illustrated in the middle row of Fig. 14.1.

14.2.3 Repeating Structure Extraction

Once the repeating segment model S is calculated, we use it to derive a repeating spectrogram model W, by taking the element-wise minimum between S and each of the r segments of the mixture spectrogram V, as exemplified in the bottom row of Fig. 14.1. The calculation of the repeating spectrogram model W is shown in Eq. 14.3.

$$\begin{aligned}
&W(i, j + (k-1)p) = \min\left\{S(i, j), V(i, j + (k-1)p)\right\} \\
&\text{for } i = 1 \ldots n, \quad j = 1 \ldots p, \quad \text{and} \quad k = 1 \ldots r
\end{aligned} \tag{14.3}$$

The idea is that, if we assume that the non-negative mixture spectrogram V is the sum of a non-negative repeating spectrogram W and a non-negative non-repeating spectrogram $V - W$, then we must have $W \leq V$, element-wise.

Once the repeating spectrogram model W is calculated, we use it to derive a soft time–frequency mask M, by normalizing W by the mixture spectrogram V, element-wise. The calculation of the soft time–frequency mask M is shown in Eq. 14.4.

$$\begin{aligned}
&M(i, j) = \frac{W(i, j)}{V(i, j)} \quad \text{with } M(i, j) \in [0, 1] \\
&\text{for } i = 1 \ldots n \quad \text{and} \quad j = 1 \ldots m
\end{aligned} \tag{14.4}$$

The rationale is that time–frequency bins that are likely to repeat at period rate p in V would have values near 1 in M and would be weighted toward the repeating background. On the other hand, time–frequency bins that are not likely to repeat at period rate p in V would have values near 0 in M and would be weighted toward the non-repeating foreground.

We could further derive a binary time–frequency mask by setting time–frequency bins in M with values above a chosen threshold $t \in [0, 1]$ to 1, while the rest is set to 0. Pilot experiments showed that the estimates sound better when using a soft time–frequency mask.

The time–frequency mask M is then symmetrized and multiplied to the STFT X of the mixture x, element-wise. The estimated background signal is obtained by inverting the resulting STFT into the time domain. The estimated foreground signal is obtained by simply subtracting the background signal from the mixture signal.

The calculation of the repeating spectrogram model W and the derivation of the soft time–frequency mask M are illustrated in the bottom row of Fig. 14.1.

Experiments on a data set of song clips showed that the original REPET can be effectively applied for music/voice separation [13, 16], performing as well as two state-of-the-art methods, one based on a pitch-based method [8] and the other based on non-negative matrix factorization (NMF) and a source-filter model [3]. Experiments showed that REPET can also be combined with other methods to improve background/foreground separation; for example, it can be used as a preprocessor to pitch detection algorithms to improve melody extraction [16], or as a postprocessor to a singing voice separation algorithm to improve music/voice separation [17].

The time complexity of the original REPET is $O(m \log m)$, where m is the number of time frames in the spectrogram. The calculation of the beat spectrum takes $O(m \log m)$, since it is based on the autocorrelation which is itself based on the fast Fourier transform (FFT), while the median filtering takes $O(m)$ (Fig. 14.2).

Figure 14.2 shows an example of music/voice separation using the orginal REPET. The mixture is a female singer (foreground) singing over a guitar accompaniment (background). The guitar has a repeating chord progression that is stable along the song. The spectrograms and the mask are shown for 5 s and up to 2.5 kHz. The file is Tamy—Que Pena Tanto Faz from the task of professionally produced music recordings of the Signal Separation Evaluation Campaign (SiSEC).[2]

The original REPET can be easily extended to handle varying repeating structures, by simply applying the method along time, on individual segments or via a sliding window (see also Sect. 14.3). For example, given a window size and an overlap percentage, the local repeating backgrounds can be successively extracted using the original REPET; the whole repeating background can then be reconstructed via overlap-add [16].

Experiments on a data set of full-track real-world songs showed that this method can be effectively applied for music/voice separation [16], performing as well as a state-of-the-art method based on NMF and a source-filter model [3]. Experiments also showed that there is a trade-off for the window size in REPET: if the window is too long, the repetitions will not be sufficiently stable; if the window is too short, there will not be sufficient repetitions [16].

[2] http://sisec.wiki.irisa.fr/tikiindex.php?page=Professionally+produced+music+recordings

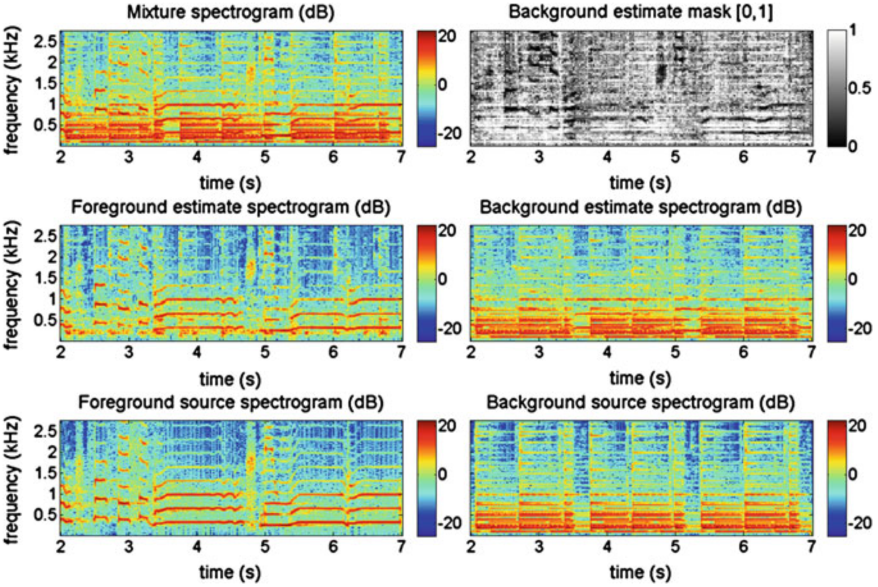

Fig. 14.2 Example of music/voice separation using the orginal REPET

14.3 Adaptive REPET

The original REPET works well when the repeating background is relatively stable (e.g., a verse or the chorus in a song); however, the repeating background can also vary over time (e.g., a verse followed by the chorus in the song). The adaptive REPET is an extension of the original REPET that can handle varying repeating structures, by estimating the time-varying repeating periods and extracting the repeating background locally, without the need for segmentation or windowing [9].

The adaptive REPET can be summarized in three stages (see Fig. 14.3): (1) identification of the repeating periods (see Sect. 14.3.1), (2) modeling of a repeating spectrogram (see Sect. 14.3.2), and (3) extraction of the repeating structure (see Sect. 14.3.3).

14.3.1 Repeating Periods Identification

The beat spectrum helps to find the global periodicity in a signal. Local periodicities can be found by computing beat spectra over successive windows. A *beat spectrogram* thus helps to visualize the variations of periodicity over time.

Given a mixture signal x, we first compute its magnitude spectrogram V (see Sect. 14.2.1). Given a window size $w \leq m$, where m is the number of time frames in

V, we then compute for every time frame j in V, the beat spectrum b_j of the local magnitude spectrogram V_j centered on j (see Sect. 14.2.1). We then concatenate the b_j's into the matrix of beat spectra B. To speed up the calculation of B, we can also use a step size s, and compute the b_j's every s frames only, and derive the rest of the values through interpolation. The calculation of B is shown in Eq. 14.5.

$$\begin{aligned}
&V_j(i,h) = V(i, h + j - \lceil\frac{w+1}{2}\rceil)\\
&A_j(i,l) = \frac{1}{w-l+1}\sum_{h=1}^{w-l+1} V_j(i,h)^2 V_j(i,h+l-1)^2 \quad \text{and}\\
&b_j(l) = \frac{1}{n}\sum_{i=1}^{n} A_j(i,l)\\
&B(l,j) = b_j(l)\\
&\text{for } i = 1\ldots n \quad \text{where } n = \frac{N}{2}+1 = \text{number of frequency channels}\\
&\text{for } h = 1\ldots w \quad \text{where } w = \text{window size}\\
&\text{for } j = 1\ldots m \quad \text{and} \quad l = 1\ldots m \quad \text{where } m = \text{number of time frames.}
\end{aligned} \tag{14.5}$$

The idea of the beat spectrogram was also introduced in [7], except that no similarity matrix is explicitly calculated here, and the dot product is used in lieu of the cosine similarity. For simplicity, we will refer to B as the beat spectrogram for the remainder of this chapter.

Once the beat spectrogram B is calculated, the first row (i.e., time lags 0) is discarded. If periodically repeating patterns are present in x, B would form horizontal lines that are periodically repeating vertically, unveiling the underlying periodically repeating structure of the mixture, as exemplified in the top row of Fig. 14.3. If variations of periodicity happen over time in x, the horizontal lines in B would show variations in their vertical periodicity.

We then use a period finder to estimate for every time frame j, the repeating period p_j from the beat spectrum b_j in B (see Sect. 14.2.1). To speed up the estimation of the p_j's, we can also use a step size s, and compute the p_j's every s frames only, and derive the rest of the values through interpolation.

The calculation of the beat spectrogram B and the estimation of the repeating periods p_j's are illustrated in the top row of Fig. 14.3.

There is no one method to compute the beat spectrum/spectrogram or to estimate the repeating period(s). We proposed to compute the beat spectrum/spectrogram using the autocorrelation and estimate the repeating period(s) using a local maximum finder (see source codes online[3]). In [9], the beat spectrogram was derived by computing the power spectrograms of the frequency channels of the power spectro-

[3] http://music.eecs.northwestern.edu/research.php?project=repet

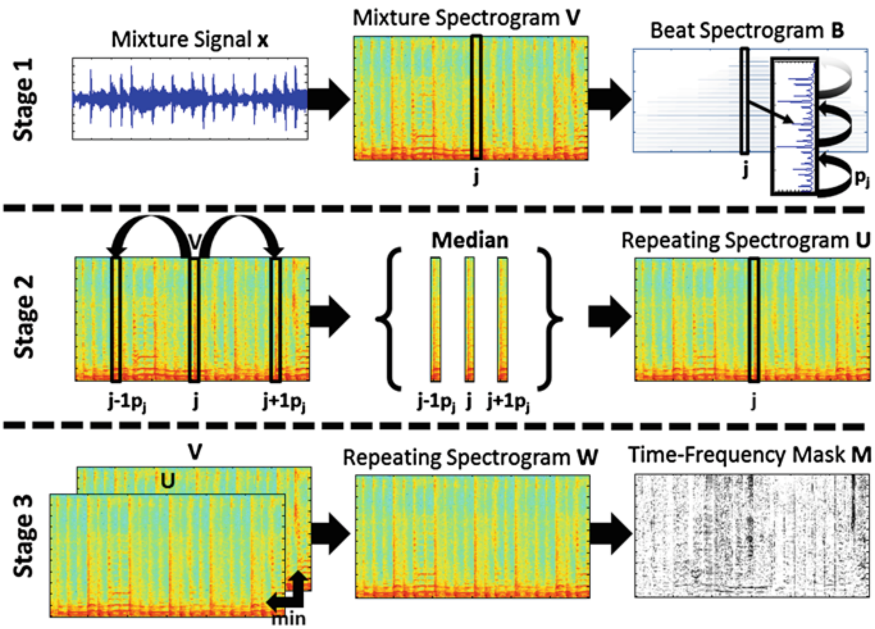

Fig. 14.3 Overview of the adaptive REPET. *Stage 1* calculation of the beat spectrogram B and estimation of the repeating periods p_j's. *Stage 2* filtering of the mixture spectrogram V and calculation of an initial repeating spectrogram model U. *Stage 3* calculation of the refined repeating spectrogram model W and derivation of the soft time–frequency mask M

gram of the mixture, and taking the element-wise mean of those power spectrograms; the repeating periods were estimated by using dynamic programming.

14.3.2 Repeating Spectrogram Modeling

Once the repeating periods p_j's are estimated, we use them to derive an initial repeating spectrogram model U. For every time frame j in the mixture spectrogram V, we derive the corresponding frame j in U by taking for every frequency channel, the median of the k frames repeating at period rate p_j around j, where k is the maximum number of repeating frames, as exemplified in the middle row of Fig. 14.3. The calculation of the initial repeating spectrogram model U is shown in Eq. 14.6.

$$\begin{aligned}
&U(i,j) = \underset{l=1\ldots k}{\text{median}}\left\{V(i, j + (l - \lceil\frac{k}{2}\rceil)p_j)\right\} \\
&\text{for } i = 1\ldots n \quad \text{and} \quad \text{for} j = 1\ldots m \qquad (14.6)\\
&\text{where } k = \text{maximum number of repeating frames} \\
&\text{where } p_j = \text{period length for frame } j.
\end{aligned}$$

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 at their period rate p_j would most likely represent repeating elements and would be captured by the median model. On the other hand, time–frequency bins with large deviations at their period rate p_j would most likely be corrupted by non-repeating elements (i.e., outliers) and would be removed by the median model.

The filtering of the mixture spectrogram V and the calculation of the initial repeating spectrogram model U are illustrated in the middle row of Fig. 14.3.

Note that, compared with the original REPET that uses the same repeating period for each time frame of the mixture spectrogram (see Sect. 14.2), the adaptive REPET uses a different repeating period for each time frame, so that it can also handle varying repeating structures where the repeating period can also change over time.

14.3.3 Repeating Structure Extraction

Once the initial repeating spectrogram model U is calculated, we use it to derive a refined repeating spectrogram model W, by taking the element-wise minimum between U and the mixture spectrogram V, as exemplified in the bottom row of Fig. 14.3. The calculation of the refined repeating spectrogram model W is shown in Eq. 14.7.

$$\begin{aligned} &W(i,j) = \min\left\{U(i,j), V(i,j)\right\} \\ &\text{for } i = 1\ldots n \quad \text{and} \quad j = 1\ldots m \end{aligned} \tag{14.7}$$

The idea is that, if we assume that the non-negative mixture spectrogram V is the sum of a non-negative repeating spectrogram W and a non-negative non-repeating spectrogram $V - W$, then we must have $W \leq V$, element-wise (see also Sect. 14.2.3).

Once the refined repeating spectrogram model W is calculated, we use it to derive a soft time–frequency mask M (see Sect. 14.2.3).

The calculation of the refined repeating spectrogram model W and the derivation of the soft time–frequency mask M are illustrated in the bottom row of Fig. 14.3.

Experiments on a data set of full-track real-world songs showed that the adaptive REPET can be effectively applied for music/voice separation [9], performing as well as a state-of-the-art method based on multiple median filtering of the mixture spectrogram at different frequency resolutions [5] (Fig. 14.4).

The time complexity of the adaptive REPET is $O(m \log m)$, where m is the number of time frames in the spectrogram. The calculation of the beat spectrogram takes $O(m \log m)$, since it is based on the beat spectrum (see Sect. 14.2.3), while the median filtering takes $O(m)$.

Figure 14.4 shows an example of music/voice separation using the adaptive REPET. The mixture is a male singer (foreground) singing over a guitar and drums

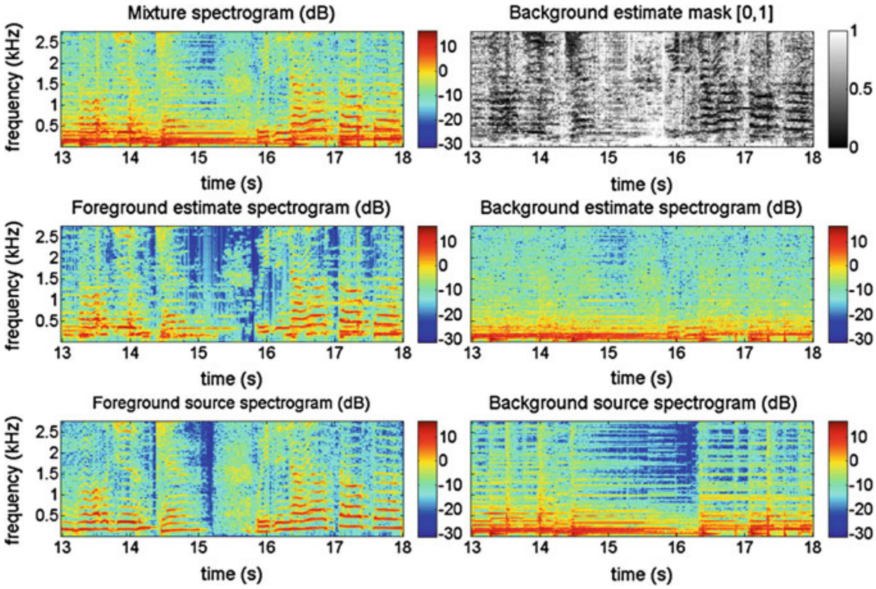

Fig. 14.4 Example of music/voice separation using the adaptive REPET

accompaniment (background). The guitar has a repeating chord progression that changes around 15 s. The spectrograms and the mask are shown for 5 s and up to 2.5 kHz. The file is Another Dreamer—The Ones We Love from the task of professionally produced music recordings of SiSEC.[4]

14.4 REPET-SIM

The REPET methods work well when the repeating background has periodically repeating patterns (e.g., jackhammer noise); however, the repeating patterns can also happen intermittently or without a global or local periodicity (e.g., frogs by a pond). REPET-SIM is a generalization of REPET that can also handle non-periodically repeating structures, by using a similarity matrix to identify the repeating elements [14, 15].

REPET-SIM can be summarized in three stages (see Fig. 14.5): (1) identification of the repeating elements (see Sect. 14.4.1), (2) modeling of a repeating spectrogram (see Sect. 14.4.2), and (3) extraction of the repeating structure (see Sect. 14.4.3).

[4] http://sisec.wiki.irisa.fr/tikiindex.php?page=Professionally+produced+music+recordings

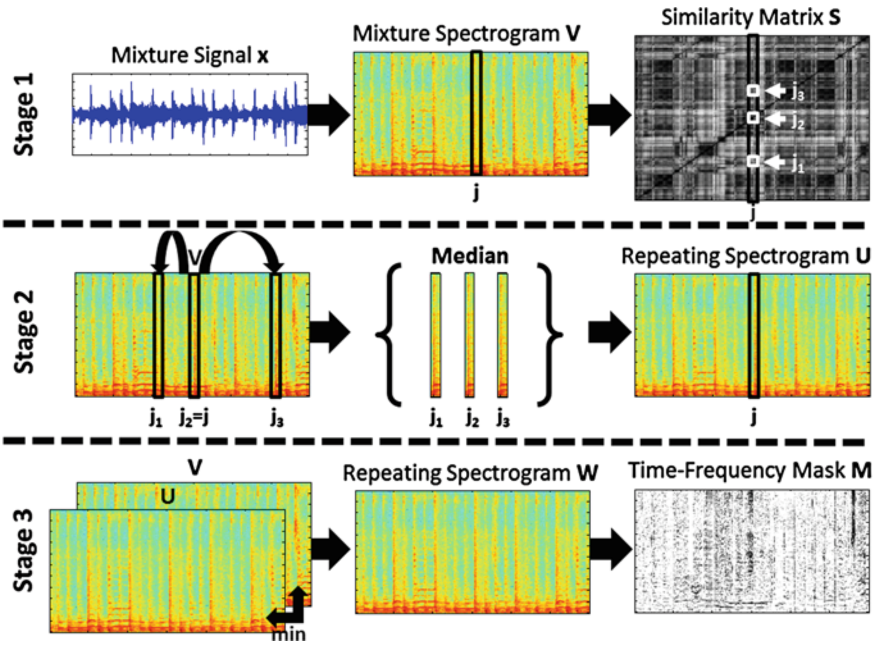

Fig. 14.5 Overview of REPET-SIM. *Stage 1* calculation of the similarity matrix S and estimation of the repeating elements j_k's. *Stage 2* filtering of the mixture spectrogram V and calculation of an initial repeating spectrogram model U. *Stage 3* calculation of the refined repeating spectrogram model W and derivation of the soft time–frequency mask M

14.4.1 Repeating Elements Identification

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

Given a mixture signal x, we first compute its magnitude spectrogram V (see Sect. 14.2.1). We then compute the similarity matrix S by multiplying transposed V and V, after normalization of the columns of V by their Euclidean norm. In other words, each point (j_a, j_b) in S measures the cosine similarity between the time frames j_a and j_b of V. The calculation of the similarity matrix S is shown in Eq. 14.8.

$$\begin{aligned} S(j_a, j_b) &= \frac{\sum_{i=1}^{n} V(i, j_a)V(i, j_b)}{\sqrt{\sum_{i=1}^{n} V(i, j_a)^2}\sqrt{\sum_{i=1}^{n} V(i, j_b)^2}} \\ \text{where } n &= \frac{N}{2} + 1 = \text{number of frequency channels} \qquad (14.8) \\ \text{for } j_a &= 1 \ldots m \quad \text{and} \quad j_b = 1 \ldots m \\ \text{where } m &= \text{number of time frames.} \end{aligned}$$

The idea of the similarity matrix was introduced in [6], 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 as the audio parametrization and the similarity measure. Pilot experiments showed that this method allows for a clearer visualization of the repeating structure in x.

Once the similarity matrix S is calculated, we use it to identify the repeating elements in the mixture spectrogram V. If repeating elements are present in x, S would form regions of high and low similarity at different times, unveiling the underlying repeating structure of the mixture, as exemplified in the top row of Fig. 14.5.

We then identify for every time frame j in V, the frames j_k's that are the most similar to frame j and save them in a vector of indices J_j. 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, the repeating elements unveiled by the similarity matrix should be those that basically compose the underlying repeating structure.

We can add the following parameters when identifying the repeating elements in the similarity matrix: t, the minimum similarity between a repeating frame and frame j; d, the minimum distance between two consecutive repeating frames; k, the maximum number of repeating frames for a frame j.

The calculation of similarity matrix S and the estimation of the repeating elements j_k's are illustrated in the top row of Fig. 14.5.

14.4.2 Repeating Spectrogram Modeling

Once the repeating elements j_k's are identified, we use them to derive an initial repeating spectrogram model 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 repeating frames j_k's given by the vector of indices J_j, as exemplified in the middle row of Fig. 14.5. The calculation of the initial repeating spectrogram model U is shown in Eq. 14.9.

$$\begin{aligned} &U(i,j) = \underset{l=1 \ldots k}{\operatorname{median}} \big\{ V(i, J_j(l) \big\} \\ &\text{where } J_j = j_1 \ldots j_k = \text{indices of repeating frames} \\ &\text{where } k = \text{maximum number of repeating frames} \\ &\text{for } i = 1 \ldots n \quad \text{and} \quad \text{for } j = 1 \ldots m. \end{aligned} \tag{14.9}$$

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 within their repeating frames j_k's would most likely represent repeating elements and would be captured by the median model. On the other hand, time–frequency bins with large deviations within their repeating frames j_k's would most likely be corrupted by non-repeating elements (i.e., outliers) and would be removed by the median model.

The filtering of the mixture spectrogram V and the calculation of the initial repeating spectrogram model U are illustrated in the middle row of Fig. 14.5.

Note that, compared with the REPET methods that look for periodically repeating elements for each time frame of the mixture spectrogram (see Sects. 14.2 and 14.3), REPET-SIM also looks for non-periodically repeating elements for each time frame, so that it can also handle non-periodically repeating structures where repeating elements can also happen intermittently.

14.4.3 Repeating Structure Extraction

Once the initial repeating spectrogram model U is calculated, we use it to derive a refined repeating spectrogram model W, as exemplified in the bottom row of Fig. 14.5 (see Sect. 14.3.3).

Once the refined repeating spectrogram model W is calculated, we use it to derive a soft time–frequency mask M (see Sect. 14.2.3).

The calculation of the refined repeating spectrogram model W and the derivation of the soft time–frequency mask M are illustrated in the bottom row of Fig. 14.5.

Experiments on a data set of full-track real-world songs showed that REPET-SIM can be effectively applied for music/voice separation [14], performing as well as a state-of-the-art method based on multiple median filtering of the mixture spectrogram at different frequency resolutions [5], and the adaptive REPET [9].

Note that FitzGerald proposed a method very similar to REPET-SIM, except that he computed a distance matrix based on the Euclidean distance and he did not use a minimum distance parameter [4].

The time complexity of the REPET-SIM is $O(m^2)$, where m is the number of time frames in the spectrogram. The calculation of the similarity matrix takes $O(m^2)$, since it is based on matrix multiplication, while the median filtering takes $O(m)$ (Fig. 14.6).

Figure 14.6 shows an example of noise/speech separation using REPET-SIM. The mixture is a female speaker (foreground) speaking in a town square (background). The square has repeating noisy elements (passers-by and cars) that happen intermittently. The spectrograms and the mask are shown for 5 s and up to 2 kHz. The file is dev_Sq1_Co_B from the task of two-channel mixtures of speech and real-world background noise of the SiSEC.[5]

REPET-SIM can be easily implemented online to handle real-time computing, particularly for real-time speech enhancement. The online REPET-SIM simply processes the time frames of the mixture one after the other given a buffer that temporally stores past frames. For every time frame being processed, the similarity

[5] http://sisec.wiki.irisa.fr/tiki-index.php?page=Two-channel+mixtures+of+speech+and+realworld+background+noise

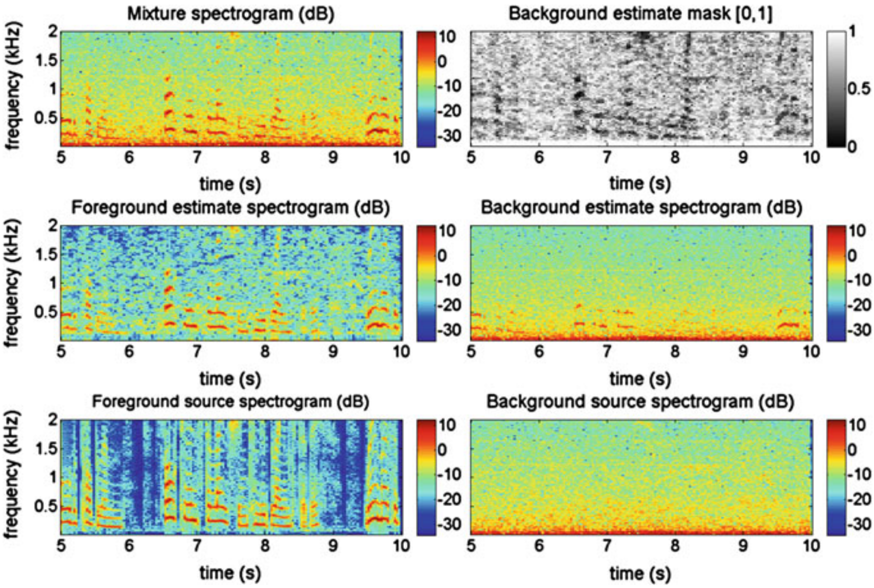

Fig. 14.6 Example of noise/speech separation using REPET-SIM

is calculated with the past frames stored in the buffer. The median is then taken between the frame being processed and its most similar frames for every frequency channel, leading to the corresponding time frame for the repeating background [15].

Experiments on a data set of two-channel mixtures of one speech source and real-world background noise showed that the online REPET-SIM can be effectively applied for real-time speech enhancement [15], performing as well as different state-of-the-art methods, one based on independent component analysis (ICA) [11], one based on the degenerate unmixing estimation technique (DUET) [20] and a minimum-statistics-based adaptive procedure [18], and one based on time differences of arrival (TDOA) and a multichannel Wiener filtering [1].

14.5 Conclusion

In this chapter, we presented *REPET*, a simple method that uses repetition as a basis for background/foreground separation in audio. In Sect. 14.2, we have presented the original REPET that aims at identifying and extracting the repeating patterns in an audio mixture, by estimating a period of the underlying repeating structure and modeling a segment of the periodically repeating background. In Sect. 14.3, we have presented the adaptive REPET, an extension of the original REPET that can directly handle varying repeating structures, by estimating the time-varying repeating periods and extracting the repeating background locally, without the need for segmentation or windowing. In Sect. 14.4, we have presented REPET-SIM, a generalization of REPET that can also handle non-periodically repeating structures, by using a similarity matrix to identify repeating elements.

Experiments on various data sets showed that REPET can be effectively applied for background/foreground separation, performing as well as different state-of-the-art approaches, while being computationally efficient. Unlike other separation approaches, REPET does not depend on special parameterizations, does not rely on complex frameworks, and does not require external information. Because it is only based on repetition, it has the advantage of being simple, fast, blind, and therefore completely and easily automatable.

More information about REPET, including source codes, audio examples, and related publications, can be found online.[6] This work was in part supported by NSF grant number IIS-0812314.

References

1. Blandin, C., Ozerov, A., Vincent, E.: Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. Signal Process. **92**(8), 1950–1960 (2012)
2. Bregman, A.S.: Auditory Scene Analysis. MIT Press, Cambridge (1990)
3. Durrieu, J.L., David, B., Richard, G.: A musically motivated mid-level representation for pitch estimation and musical audio source separation. IEEE J. Sel. Top. Sig. Process. **5**(6), 1180–1191 (2011)
4. FitzGerald, D.: Vocal separation using nearest neighbours and median filtering. In: 23rd IET Irish Signals and Systems Conference. Maynooth, Ireland (2012)
5. FitzGerald, D., Gainza, M.: Single channel vocal separation using median filtering and factorisation techniques. ISAST Trans. Electron. Signal Process. **4**(1), 62–73 (2010)
6. Foote, J.: Visualizing music and audio using self-similarity. In: 7th ACM International Conference on Multimedia, pp. 77–80. Orlando, FL, USA (1999)
7. Foote, J., Uchihashi, S.: The beat spectrum: a new approach to rhythm analysis. In: IEEE International Conference on Multimedia and Expo, pp. 881–884. Tokyo, Japan (2001)
8. Hsu, C.L., Jang, J.S.R.: On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. IEEE Trans Audio Speech Lang. Process. **18**(2), 310–319 (2010)
9. Liutkus, A., Rafii, Z., Badeau, R., Pardo, B., Richard, G.: Adaptive filtering for music/voice separation exploiting the repeating musical structure. In: 37th International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan (2012)
10. McDermott, J.H., Wrobleski, D., Oxenham, A.J.: Recovering sound sources from embedded repetition. Proc Nat. Acad. Sci. U.S.A. **108**(3), 1188–1193 (2011)
11. Nesta, F., Matassoni, M.: Robust automatic speech recognition through on-line semi blind source extraction. In: CHIME 2011 Workshop on Machine Listening in Multisource Environments, pp. 18–23. Florence, Italy (2011)
12. Piccardi, M.: Background subtraction techniques: a review. In: IEEE International Conference on Systems, Man and Cybernetics. The Hague, The Netherlands (2004)
13. Rafii, Z., Pardo, B.: A simple music/voice separation system based on the extraction of the repeating musical structure. In: 36th International Conference on Acoustics, Speech and Signal Processing. Prague, Czech Republic (2011)

[6] http://music.eecs.northwestern.edu/research.php?project=repet

14. Rafii, Z., Pardo, B.: Music/voice separation using the similarity matrix. In: 13th International Society for Music Information Retrieval. Porto, Portugal (2012)
15. Rafii, Z., Pardo, B.: Online REPET-SIM for real-time speech enhancement. In: 38th International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada (2013)
16. Rafii, Z., Pardo, B.: REpeating Pattern Extraction Technique (REPET): A simple method for music/voice separation. IEEE Trans. Audio Speech Lang. Process **21**(1), 71–82 (2013)
17. Rafii, Z., Sun, D.L., Germain, F.G., Mysore, G.J.: Combining modeling of singing voice and background music for automatic separation of musical mixtures. In: 14th International Society for Music Information Retrieval. Curitiba, PR, Brazil (2013).
18. Rangachari, S., Loizou, P.C.: A noise-estimation algorithm for highly non-stationary environments. Speech Commun. **48**(2), 220–231 (2006)
19. Rubin, E.: Synsoplevede Figurer. Gyldendal, Skive (1915)
20. Özgür Yilmaz, Rickard, S.: Blind separation of speech mixtures via time–frequency masking. IEEE Trans. Signal Process. **52**(7), 1830–1847 (2004)