MUSIC/VOICE SEPARATION USING THE SIMILARITY MATRIX

Zafar RAFII
Northwestern University
EECS Department
Evanston, IL, USA
zafarrafii@u.northwestern.edu

Bryan PARDO
Northwestern University
EECS Department
Evanston, IL, USA
pardo@northwestern.edu

ABSTRACT

Repetition is a fundamental element in generating and perceiving structure in music. Recent work has applied this principle to separate the musical background from the vocal foreground in a mixture, by simply extracting the underlying repeating structure. While existing methods are effective, they depend on an assumption of periodically repeating patterns. In this work, we generalize the repetition-based source separation approach to handle cases where repetitions also happen intermittently or without a fixed period, thus allowing the processing of music pieces with fast-varying repeating structures and isolated repeating elements. Instead of looking for periodicities, the proposed method uses a similarity matrix to identify the repeating elements. It then calculates a repeating spectrogram model using the median and extracts the repeating patterns using a time-frequency masking. Evaluation on a data set of 14 full-track real-world pop songs showed that use of a similarity matrix can overall improve on the separation performance compared with a previous repetition-based source separation method, and a recent competitive music/voice separation method, while still being computationally efficient.

1. INTRODUCTION

A system that can efficiently separate a song into foreground (e.g. the soloist or voice) and background (the musical accompaniment) components would be of great interest for a wide range of applications. These applications include instrument/vocalist identification, music/voice transcription, melody extraction, audio remixing, and karaoke.

While there are many approaches that have been applied to this problem (see Section 2), one promising approach is to use analysis of the repeating structure in the audio. Many musical pieces are characterized by an underlying repeating structure (e.g. drum loop or 4-measure vamp) over which varying elements are superimposed. This is especially true for pop songs where a singer often overlays varying vocals on a repeating accompaniment.

Recent work (see Section 2) has exploited repetition to separate the repeating musical background from the non-repeating vocal foreground. This work has relied on the assumption that there is a global or a local period of repetition in the musical background. Should repeated elements be present (e.g. reuse of the same chord voicing in the piano) but performed in a way that is not obviously periodic (e.g. occasional chordal piano "fills" at the appropriate moments), existing repetition-based approaches fail.

In this work, we generalize the repetition-based source separation approach to handle cases where repetitions also happen intermittently or without a fixed period. Instead of looking for periodicities, the proposed method identifies repeating elements by looking for similarities, by means of a similarity matrix. Once identified, median filtering is then performed on the repeating elements to calculate a repeating spectrogram model for the background. A time-frequency mask can finally be derived to extract the repeating patterns (see Section 3). This allows the processing of music pieces with fast-varying repeating structures and isolated repeating elements, without the need to identify periods of the repeating structure beforehand.

The rest of this paper is organized as follows. Section 2 describes the related work. Section 3 introduces the proposed method. Section 4 presents an evaluation of the proposed method on a data set of 14 full-track real-world pop songs, against a previous repetition-based source separation method, and a recent competitive music/voice separation method. Section 5 concludes this article.

2. RELATED WORK

There have been a number of approaches applied to the problem of separating the foreground (typically the voice) from the background. In stereo recordings, panning information (e.g. the vocalist is typically panned to the middle) can be applied. Such approach fails when the vocalist is not center-panned (e.g. many Beatles recordings). Cross-channel timing and amplitude differences can be applied in more complex frameworks, such as in the Degenerate Unmixing Estimation Technique (DUET) [8]. This approach is difficult to apply to pop music, due to the reverberant effects added, as well as the violation of the sparsity assumption for music mixtures.

Other music/voice separation methods focus on mod-

eling either the music signal, by generally training an accompaniment model from the non-vocal segments [6, 11], or the vocal signal, by generally extracting the predominant pitch contour [7, 9], or both signals via hybrid models [1, 13]. Most of these methods require a training phase on audio with labeled vocal/non-vocal segments.

Recently, a relatively simple approach has also been proposed for music/voice separation. The method is based on a median filtering of the mixture spectrogram at different frequency resolutions, in such a way that the harmonic and percussive elements of the accompaniment can be smoothed out, leaving out the vocals [3].

Another recent and promising approach is to apply analysis of the repeating structure in the audio to extract the repeating musical background from the non-repeating vocal foreground. In this work, we focus on this approach.

The first method to explicitly use repetition to separate the musical background from the vocal foreground is the REpeating Pattern Extraction Technique (REPET) [12]. The method seeks to identify a global period for the repeating structure, so that it can build a model of the repeating background. This model is then used to construct a time-frequency mask to separate the repeating musical background from the non-repeating vocal foreground.

The original REPET method can be successfully applied for music/voice separation on short excerpts (e.g. 10 second verse) [12]. For complete music pieces, the repeating background is likely to vary over time (e.g. verse followed by chorus). An extended version of REPET was therefore later introduced to handle variations in the repeating structure [10]. Rather than finding a global period, the method tracks local periods of the repeating structure. In both cases, the algorithm needs to identify periods of the repeating structure, as both methods assume periodically repeating patterns.

In this work, we propose to generalize the repetition-based source separation approach to handle cases where repetitions also happen intermittently or without a fixed period, by using a similarity matrix.

3. PROPOSED METHOD

3.1 Similarity Matrix

The similarity matrix 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. Since repetition/similarity is what makes the structure in music, a similarity matrix calculated from an audio signal can help to reveal the musical structure that underlies it [4].

Given a single-channel mixture signal x, we first calculate its Short-Time Fourier Transform (STFT) X, using half-overlapping Hamming windows of N samples length. We then derive the magnitude spectrogram V by taking the absolute value of the elements of X, after discarding the symmetric part, while keeping the DC component.

We then define the similarity matrix S as the matrix multiplication between 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 the mixture spectrogram V. The calculation of the similarity matrix S is shown in Equation 1.

$$S(j_a, j_b) = \frac{\sum_{i=1}^{n} V(i, j_a)V(i, j_b)}{\sqrt{\sum_{i=1}^{n} V(i, j_a)^2}\sqrt{\sum_{i=1}^{n} V(i, j_b)^2}} \quad (1)$$
$$\text{where } n = N/2 + 1 = \text{\# frequency channels}$$
$$\forall j_a, j_b \in [1, m] \text{ where } m = \text{\# time frames}$$

3.2 Repeating Elements

Once the similarity matrix S is calculated, we use it to identify the repeating elements in the mixture spectrogram V. For all the frames j in V, we look for the frames that are the most similar to the given frame j and save them in a vector of indices J_j. Assuming that the non-repeating foreground ($\approx$ voice) is sparse and varied compared to the repeating background ($\approx$ music) - a reasonable assumption for voice in music, the repeating elements revealed by the similarity matrix should be those that form the underlying repeating structure ($\approx$ music). The use of a similarity matrix actually allows us to identify repeating elements that do not necessarily happen in a periodic fashion.

We add the following constraint parameters to the algorithm. To limit the number of repeating frames considered similar to the given frame j, we define k, the maximum allowed number of repeating frames. We define t, the minimum allowed threshold for the similarity between a repeating frame and the given frame ($t \in [0, 1]$). Consecutive frames can exhibit high similarity without representing new instances of the same structural element, since frame duration is unrelated to the duration of musical elements. We therefore define d, the minimum allowed (time) distance between two consecutive repeating frames deemed to be similar enough to indicate a repeating element.

3.3 Repeating Model

Once the repeating elements have been identified for all the frames j in the mixture spectrogram V through their corresponding vectors of indices J_j, we use them to derive a repeating spectrogram model W for the background. For all the frames j in V, we derive the corresponding frame j in W by taking the median of the corresponding repeating frames whose indices are given by vector J_j, for every frequency channel. The calculation of the repeating spectrogram model W is shown in Equation 2.

$$W(i, j) = \underset{l \in [1,k]}{median}\{V(i, J_j(l)\}$$
$$\text{where } J_j = [j_1 \dots j_k] = \text{indices of repeating frames}$$
$$\text{where } k = \text{maximum number of repeating frames} \quad (2)$$
$$\forall i \in [1, n] = \text{frequency channel index}$$
$$\forall j \in [1, m] = \text{time frame index}$$

The rationale is that, assuming that the non-repeating foreground ($\approx$ voice) has a sparse time-frequency representation compare to the time-frequency representation of the repeating background ($\approx$ music), time-frequency bins

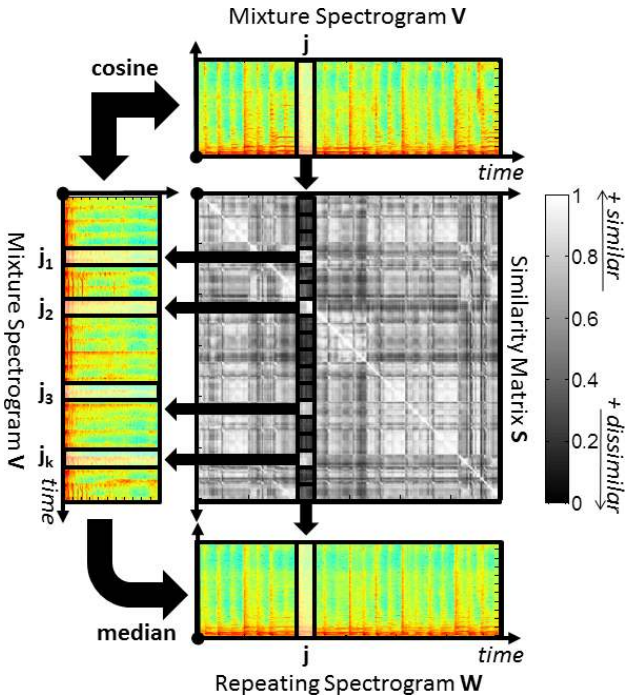

Figure 1. Derivation of the repeating spectrogram model W: (1) compute the similarity matrix S from the mixture spectrogram V using the cosine similarity measure; (2) for all frames j in V, identify the k frames $j_1 \dots j_k$ that are the most similar to frame j using S; (3) derive frame j of the repeating spectrogram model W by taking the median of the k frames $j_1 \dots j_k$ of V, for every frequency channel.

with little deviations between repeating frames would constitute a repeating pattern and would be captured by the median. Accordingly, time-frequency bins with large deviations between repeating frames would constitute a non-repeating pattern and would be removed by the median. The derivation of the repeating spectrogram model W from the mixture spectrogram V using the similarity matrix S is illustrated in Figure 1.

3.4 Time-frequency Mask

Once the repeating spectrogram model W is calculated, we use it to derive a time-frequency mask M. But first, we need to create a refined repeating spectrogram model W' for the background, by taking the minimum between W and V, for every time-frequency bin. Indeed, as noted in [10], 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 time-frequency bins in W can at most have the same value as the corresponding time-frequency bins in V. In other words, we want $W \leq V$, for every time-frequency bin; hence the use of the minimum function.

We then derive a time-frequency mask M by normalizing W' by V, for every time-frequency bin. The rationale is that time-frequency bins that are likely to constitute a repeating pattern in V will have values near 1 in M and will be weighted toward the repeating background ($\approx$ music). Accordingly, time-frequency bins that are unlikely to constitute a repeating pattern in V will have values near 0 in M and will be weighted toward the non-repeating foreground ($\approx$ voice). The calculation of the time-frequency mask M is shown in Equation 3.

$$\begin{aligned} W'(i,j) &= \min\big(W(i,j), V(i,j)\big) \\ M(i,j) &= \frac{W'(i,j)}{V(i,j)} \quad \text{with } M(i,j) \in [0,1] \\ \forall i \in [1,n] &= \text{frequency channel index} \\ \forall j \in [1,m] &= \text{time frame index} \end{aligned} \tag{3}$$

The time-frequency mask M is then symmetrized and applied to the STFT X of the mixture signal x. The estimated music signal is finally obtained by inverting the resulting STFT into the time domain. The estimated voice signal is obtained by simply subtracting the music signal from the mixture signal.

4. EVALUATION

4.1 Competitive Methods & Data Set

We label the proposed method, based on the use of a similarity matrix, *Proposed*. We compare separation performance of *Proposed* with two competitive music/voice separation methods on a data set of 14 full-track pop songs.

The first competitive method is an extension of the original REPET algorithm to handle variations in the underlying repeating structure [10]. We refer to this method as *REPET+*. The method first tracks local periods of the underlying repeating structure using a beat spectrogram, then models local estimates of the repeating background using the median, and finally extracts the repeating patterns from the mixture using a time-frequency mask. For the comparison, we used the separation results of *REPET+* with soft time-frequency masking and high-pass filtering with a cutoff frequency of 100 Hz on the voice estimates, as published in [10].

The second competitive method is the Multipass Median Filtering-based Separation (MMFS), another recently proposed simple and novel approach for music/voice separation [3]. The method is based on a median filtering of the mixture spectrogram at different frequency resolutions, in such a way that the harmonic and percussive elements of the accompaniment can be smoothed out, leaving out the vocals. For the comparison, we used the separation results of the best version of MMFS out of the four proposed versions, with high-pass filtering with a cutoff frequency of 100 Hz on the voice estimates, as published in [3].

The data set consists of 14 full-track real-world pop songs, in the form of split stereo WAVE files sampled at 44.1 kHz, with the accompaniment and vocals on the left and right channels, respectively. These 14 stereo sources were created from recordings released by the band *The Beach Boys*, where some of the accompaniments and vocals were made available as split stereo tracks [1] and separated tracks [2]. This data set was used in [10] for the evalu-

[1] Good Vibrations: Thirty Years of The Beach Boys, 1993
[2] The Pet Sounds Sessions, 1997

ation of *REPET+* against the best version of MMFS.

Following the framework adopted in [3] and [10], we then used those 14 stereo sources to create three data sets of 14 mono mixtures, by mixing the channels at three different voice-to-music ratios: -6 dB (music is louder), 0 dB (same original level), and 6 dB (voice is louder). Note that we are using the exact same data set as in [10], however it is not the exact same data set that was used in [3]. The authors of [3] did not mention which tracks they used for their experiments and also unlike them, but as in [10], we process the full tracks without segmenting them beforehand, since *Proposed* can handle long recordings, and this without memory or computational constraints.

4.2 Algorithm Parameters & Separation Measures

We calculated the STFT of each mixture for each of the three mixture sets (-6, 0, and 6 dB) using half-overlapping Hamming windows of $N = 2048$ samples length, corresponding to a duration of 46.4 milliseconds at a sampling frequency of 44.1 kHz. We then processed each mixture using *Proposed*. The parameters were fixed as follows: maximum number of repeating frames $k = 100$, minimum threshold for the similarity between a repeating frame and the given frame $t = 0$, and minimum distance between two consecutive repeating frames $d = 1$ second. Pilot experiments showed that those parameters lead to overall good separation results. For the comparison, we also applied a high-pass filtering with a cutoff frequency of 100 Hz on the voice estimates. This means that all the energy under 100 Hz in the voice estimates is transferred to the corresponding music estimates. The rational is that singing voice rarely happen below 100 Hz.

We measured separation performance by employing the *BSS_EVAL toolbox* [2]. The toolbox proposes a set of now widely adopted measures that intend to quantify the quality of the separation between a source and its corresponding estimate: Source-to-Distortion Ratio (SDR), Sources-to-Interferences Ratio (SIR), and Sources-to-Artifacts Ratio (SAR). Following the framework adopted in [3] and [10], we measured SDR, SIR, and SAR on segments of 45 second length from the music and voice estimates. Higher values of SDR, SIR, and SAR suggest better separation performance. We chose to use those measures because they are widely known and used, and also because they have been shown to be well correlated with human assessments of signal quality [5].

4.3 Comparative Results & Statistical Analysis

Figures 2, 3, and 4 show the separation performance using the SDR, SIR, and SAR, respectively, in dB, for the music component (top row) and the voice component (bottom row), at voice-to-music mixing ratio of -6 dB (left column), 0 dB (middle column), and 6 dB (right column). In each column, from left to right, the first results correspond to the best version of MMFS (*MMFS*), where the means of the distributions are represented as crosses (the standard deviations were not reported in [3]). The second results correspond to the extension of REPET for varying repeating structures (*REPET+*), where the means and standard deviations of the distributions are represented as error bars. The third results correspond to the proposed method with similarity matrix (*Proposed*), where the means and standard deviations of the distributions are represented as error bars. The mean values are displayed. Higher values are better.

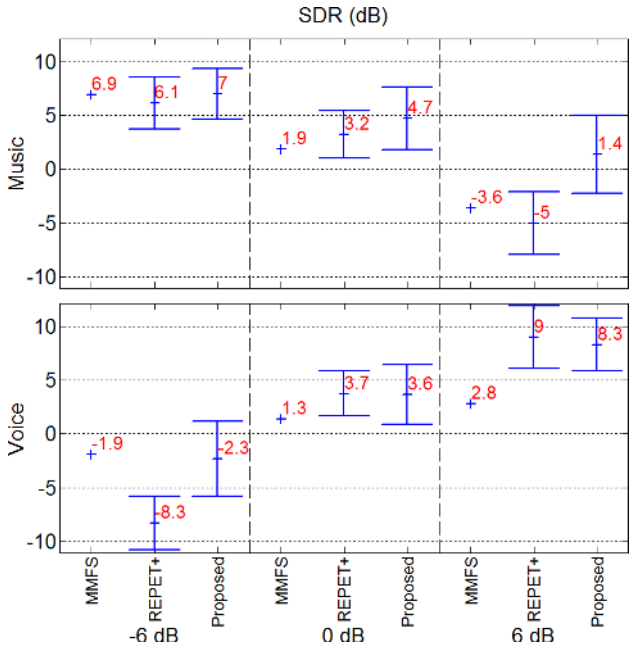

Figure 2. Separation performance using the SDR in dB, for the music component (top row) and the voice component (bottom row), at voice-to-music mixing ratio of -6 dB (left column), 0 dB (middle column), and 6 dB (right column), using the best version of MMFS (*MMFS*) (means represented as crosses), the extension of REPET for varying repeating structures (*REPET+*), and the proposed method with similarity matrix (*Proposed*) (means and standard deviations represented as error bars). Mean values are displayed. Higher values are better.

We compared the three different methods including a high-pass filtering with a cutoff frequency of 100 Hz on the voice estimates, because such post-processing of the estimates typically helps to produce better separation results. For our proposed method, the high-pass filtering increased SDR and SIR, for both the music and voice estimates. SAR however increased only for the music estimates. This is probably due to the fact that, although improving the separation performance overall, using a high-pass filtering on the voice estimates creates "holes" in their time-frequency representation, which tend to increase the separation artifacts, hence the decrease of SAR for the voice estimates.

As we can see in Figures 2, 3, and 4, as the voice-to-music ratio gets larger, SDR, SAR, and SIR get lower for the music estimates and larger for the voice estimates, and vice versa. This is an intuitive result also observed for *MMFS* and *REPET+*. Indeed, as the voice component gets louder compared to the music component, it then

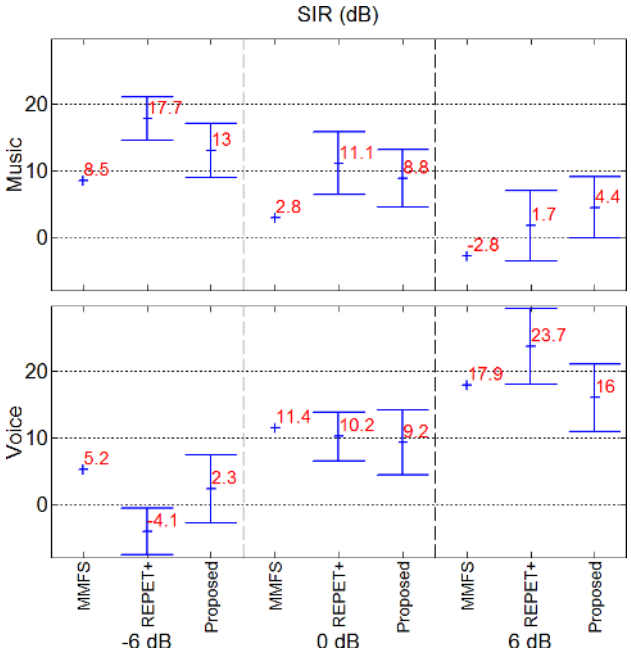

Figure 3. Separation performance using the SIR in dB.

becomes easier to extract the voice component, and accordingly harder to extract the music component, and vice versa. A multiple comparison test showed that those results were statistically significant in each case. We used an Analysis of Variance (ANOVA) when the compared distributions were all normal, and a Kruskal-Wallis test when at least one of the compared distributions was not normal. We used a Jarque-Bera normality test to determine if a distribution was normal or not.

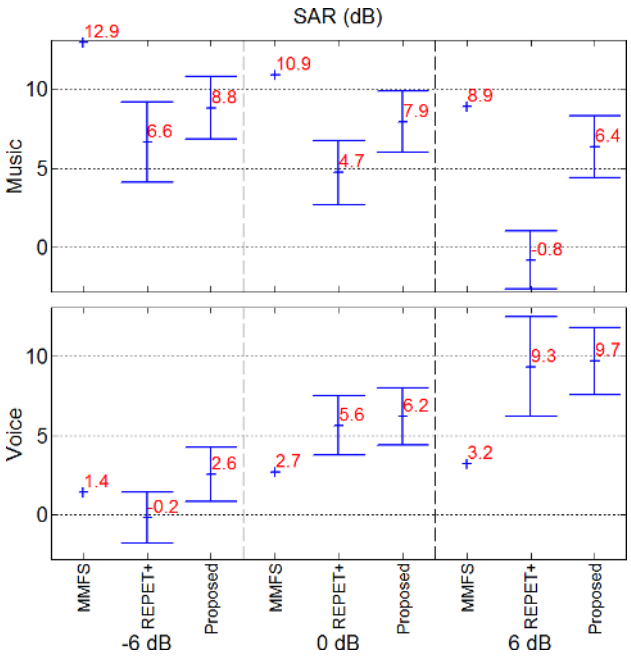

Figure 4. Separation performance using the SAR in dB.

As we can see in Figures 2, 3, and 4, compared with *MMFS*, *Proposed* gave overall better SDR for both the music and voice estimates, better SIR for the music estimates, and better SAR for the voice estimates, and this for all the three voice-to-music ratios. A one-sample t-test comparing the means of the distributions of *Proposed* with the means of *MMFS* (since the only values provided in [3] were the means) showed that those results were statistically significant in each case, except for the SDR at voice-to-music ratio of -6 dB, where the improvement of *Proposed* compared with *MMFS* was not significant for the music estimates, and a decrease, although not significant, was observed for the voice estimates. This suggest that, compared with *MMFS*, *Proposed* has globally better separation performance, particularly it is better at removing the "vocal" interferences from the accompaniment, and at limiting the separation artifacts in the voice estimates.

As we can also see in Figures 2, 3, and 4, for the music estimates, compared with *REPET+*, *Proposed* gave overall better SDR and SAR for all the three voice-to-music ratios, and better SIR at voice-to-music ratio of 6 dB. A multiple comparison test showed that those results were statistically significant in each case, except for the SDR at voice-to-music ratio of -6 dB where the improvement of *Proposed* compared with *REPET+* was not significant. For the voice estimates, compared with *REPET+*, *Proposed* gave overall better SAR for all the three voice-to-music ratios, and better SDR and SIR at voice-to-music ratio of -6 dB. A multiple comparison test showed that those results were statistically significant in each case, except for the SAR where the improvement of *Proposed* compared with *REPET+* was only significant at voice-to-music ratio of -6 dB. We used ANOVA when the compared distributions were all normal, and a Kruskal-Wallis test when at least one of the compared distributions was not normal. This suggest that, compared with *REPET+*, *Proposed* has globally better separation performance for a component (music or voice), as the given component becomes softer compared with the other one.

The average computation time of *Proposed* over all the mixtures and all of the three mixture sets (-6, 0, and 6 dB) was 0.563 second for 1 second of mixture, when implemented in Matlab on a PC with Intel(R) Core(TM)2 Quad CPU of 2.66 GHz and 6 GB of RAM. In other words, *Proposed* can perform music/voice separation of a mixture audio in half the time of the playback of the audio, for recordings of the length of a typical pop song. This is encouraging, since *Proposed* builds a similarity matrix that is $O(n^2)$, where n is the length of the audio file. As a point of comparison, the average computation time for *REPET+* for the exact same data set was 1.1830 second for 1 second of mixture [10].

5. CONCLUSION

In this work, we have proposed a generalization of the REpeating Pattern Extraction Technique (REPET) method for the task of music/voice separation, based on the calculation of a similarity matrix. The REPET approach is based on the separation of a musical background from a vocal foreground, by extraction of the underlying repeating structure. The basic idea is to identify elements that exhibit similar-

ity, and compare them to repeating models derived from them to extract the repeating patterns.

Unlike the previous REPET methods that assume periodically repeating patterns, the proposed method with similarity matrix generalizes to repeating structures where repetitions can also happen intermittently or without a fixed period, therefore allowing the processing of music pieces with fast-varying repeating structures and isolated repeating elements, without the need to identify periods of the underlying repeating structure beforehand.

Evaluation on a data set of 14 full-track real-world pop songs showed that the proposed generalization of REPET with similarity matrix can overall improve on the separation performance compared with the extension of REPET for varying repeating structures, and another recent competitive music/voice separation method based on median filtering, while still being computationally efficient. Given the SDR, which can be understood as a measure of the overall quality of the separation, our evaluation showed that when the results between the proposed method and the competitive methods were statistically significant, the proposed method gave higher results, and this compared with both the competitive methods.

The proposed generalization of REPET is only based on a similarity matrix. In other words, it does not depend on particular features, does not rely on complex frameworks, and does not need prior training. Because it is only based on self-similarity, it has the advantage of being simple, fast, blind, and therefore completely and easily automatable.

6. ACKNOWLEDGMENTS

The authors would like to thank Antoine Liutkus, Roland Badeau, Gaël Richard, and their colleagues from Telecom ParisTech for the fruitful discussions. This work was supported by NSF grant number IIS-0812314.

7. REFERENCES

[1] Jean-Louis Durrieu, Bertrand David, and Gaël Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. *IEEE Journal on Selected Topics on Signal Processing*, 5(6):1180–1191, October 2011.

[2] Cedric Févotte, Rémi Gribonval, and Emmanuel Vincent. BSS_EVAL toolbox user guide. Technical Report 1706, IRISA, Rennes, France, April 2005. http://www.irisa.fr/metiss/bss eval/.

[3] Derry FitzGerald and Mikel Gainza. Single channel vocal separation using median filtering and factorisation techniques. *ISAST Transactions on Electronic and Signal Processing*, 4(1):62–73, 2010.

[4] Jonathan Foote. Visualizing music and audio using self-similarity. In *7th ACM International Conference on Multimedia (Part 1)*, pages 77–80, Orlando, FL, USA, October 30-November 0 1999.

[5] Brendan Fox, Andrew Sabin, Bryan Pardo, and Alec Zopf. Modeling perceptual similarity of audio signals for blind source separation evaluation. In *7th International Conference on Independent Component Analysis*, pages 454–461, London, UK, September 09-12 2007.

[6] Jinyu Han and Ching-Wei Chen. Improving melody extraction using probabilistic latent component analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 22-27 2011.

[7] Chao-Ling Hsu and Jyh-Shing Roger Jang. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):310–319, February 2010.

[8] Alexander Jourjine, Scott Rickard, and Özgür Yilmaz. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 2985–2988, Istanbul, Turkey, June 5-9 2000.

[9] Yipeng Li and DeLiang Wang. Separation of singing voice from music accompaniment for monaural recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1475–1487, May 2007.

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

[11] Alexey Ozerov, Pierrick Philippe, Frédéric Bimbot, and Rémi Gribonval. Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1564–1578, July 2007.

[12] Zafar Rafii and Bryan Pardo. A simple music/voice separation system based on the extraction of the repeating musical structure. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 22-27 2011.

[13] Tuomas Virtanen, Annamaria Mesaros, and Matti Ryynänen. Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. In *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, pages 17–20, Brisbane, Australia, 21 September 2008.