ADAPTIVE FILTERING FOR MUSIC/VOICE SEPARATION EXPLOITING THE REPEATING MUSICAL STRUCTURE

Antoine Liutkus[⋆] *Zafar Rafii*[†] *Roland Badeau*[⋆] *Bryan Pardo*[†] *Gaël Richard*[⋆]

[⋆] Institut Telecom, Telecom ParisTech, CNRS LTCI, France.
[†]Northwestern University, EECS Department, Evanston, IL, USA

ABSTRACT

The separation of the lead vocals from the background accompaniment in audio recordings is a challenging task. Recently, an efficient method called REPET (REpeating Pattern Extraction Technique) has been proposed to extract the repeating background from the non-repeating foreground. While effective on individual sections of a song, REPET does not allow for variations in the background (e.g. verse vs. chorus), and is thus limited to short excerpts only. We overcome this limitation and generalize REPET to permit the processing of complete musical tracks. The proposed algorithm tracks the period of the repeating structure and computes local estimates of the background pattern. Separation is performed by soft time-frequency masking, based on the deviation between the current observation and the estimated background pattern. Evaluation on a dataset of 14 complete tracks shows that this method can perform at least as well as a recent competitive music/voice separation method, while being computationally efficient.

Index Terms— Music/voice separation, repeating pattern, time-frequency masking, adaptive algorithms

1. INTRODUCTION

This work focuses on separating the singing voice signal from its musical background in audio excerpts. This is a special case of separating out a human voice from structured background noise (e.g. music, hammering, engine noise). This highly challenging task has important practical applications, such as melody transcription from musical mixtures (making music audio databases searchable by sung melodies), removal of repetitive background noise for improved speech recognition, automatic karaoke and, more generally, *active listening* scenarios, that are defined as the ability for the user to directly interact with the musical content of the tracks.

Current trends in audio source separation are based on a *filtering* paradigm, in which the sources are recovered through the direct processing of the mixtures. When considering Time-Frequency (TF) representations, this filtering can be approximated as an element-wise weighting of the TF representations (e.g. Short-Time Fourier Transform) of the mixtures. When individual TF bins are assigned weights of either 0 (e.g. background) or 1 (e.g. foreground), this is known as binary TF masking [14]. In this case, the energy from each TF bin is assigned to just one source (foreground or background). On the other hand, allowing values between 0 and 1 permits to assign energy proportionally to each source. This is known as a soft weighting strategy [1, 2]. The point of such methods is to estimate a TF mask to apply to the mixtures and separate sources.

This work is partly funded by the Quaero Programme, by OSEO, French State Agency for Innovation, and by NSF grant number IIS-0812314.

Typical music/voice separation methods focus on modeling either the music signal, by generally training an accompaniment model from the non-vocal segments [8, 12], or the vocal signal, by generally extracting the predominant pitch contour [10, 9], or both signals via hybrid models [15, 3]. Most of those methods need to identify the vocal segments beforehand, typically using audio features such as the Mel-Frequency Cepstrum Coefficients (MFCC). Among those methods, works such as [12, 3] model each source of interest as the sum of locally stationary signals, characterized by constant normalized power spectra and time-varying energy. The estimation of the parameters of such models is done using tensor factorizations [5, 11] and separation is then consistently performed through the use of an adaptive Wiener-like filter [1, 2, 11].

Another path of research exploits the fact that a binary mask can be understood as a classification problem where each TF bin is either associated to the voice or to the music signal. If a model of the voice is available, then TF bins can be classified as belonging to the music if the corresponding observations are far from the model, thus defining a binary mask. With this in mind, a recently proposed technique called REPET (REpeating Pattern Extraction Technique) focuses on modeling the *accompaniment* instead of the vocals [13]. REPET starts from the observation that many popular music recordings can be understood as a *repeating* musical background, over which a voice signal is superimposed that does not exhibit any immediate repeating structure. Based on this observation, a model for the background signal can be computed, provided its period is correctly estimated. This technique proved to be highly effective for excerpts with a relatively stable repeating background (e.g. 10 second verse). For longer musical excerpts however, the musical background is likely to vary over time (e.g. verse followed by chorus), limiting the length of excerpt that REPET can be applied to. Furthermore, the binary TF masking used in REPET leads to noise artifacts.

In this work, we extend REPET to the case where the background is locally periodic, thus allowing the processing of long musical signals. Variations in the repeating background (e.g. verse vs. chorus) can then be handled without the need of a prior segmentation of the audio (e.g. verse/chorus/verse). We also present a soft-masking strategy, where the TF mask is not binary anymore. Such an extension of REPET involves three main challenges. First, it relies on the estimation of the time-varying period of the repeating background. Second, it requires estimating the corresponding locally-periodic musical signal. Third, using this estimate, it involves the derivation of a TF mask to perform separation.

The article is organized as follows. First, we present the framework we use for modeling the background signal in section 2, along with the corresponding method for separation. In section 3, we focus on how to estimate the time-varying period of the background and its power spectrogram. Finally, we present an evaluation of the proposed method in section 4.

2. MODEL

2.1. Notations

Let $\{x_n\}_{n=1\cdots N}$ denote a discrete-time *mixture* signal of length N, which is the sum of two signals: the lead (voice) signal $\{v_n\}_{n=1\cdots N}$ and the background signal $\{b_n\}_{n=1\cdots N}$. Let us call $\mathcal{F}\{x\}$ the Short-Time Fourier Transform (STFT) of x. Let X, V, and $B \in \mathbb{R}_+^{F\times T}$ be the power spectrograms (defined as the squared magnitude of the STFT) of x, v and b, respectively. F is the number of *frequency channels* and T the number of *time frames*. In this study, we only consider mono recordings, since the proposed technique can be applied separately on the left and right channels of stereo mixtures.

If the signals are modeled as locally stationary Gaussian processes, it can be shown [1, 11] that an estimate $\hat{b}$ of the background is given as an adaptive Wiener-like filtering of the mixture, i.e.:

$$\hat{b} = \mathcal{F}^{-1}\{W_b \cdot \mathcal{F}\{x\}\} \quad (1)$$

where $\cdot$ stands for the component-wise multiplication and where W_b is called a TF *mask*. W_b is such that for each TF bin (f,t), $W_b(f,t) \in [0\,1]$ and can be understood as the probability that the energy in bin (t,f) comes from source b. Likewise, an estimate $\hat{v}$ for v is given as: $\hat{v} = \mathcal{F}^{-1}\{(1-W_b)\cdot\mathcal{F}\{x\}\}$.

2.2. Repeating Patterns

The background signal b is assumed to be *locally spectrally-periodic* with a typical repetition period $\in [1\text{s}\,5\text{s}]$. We define a spectrally-periodic signal of period T_0 as a signal such that each frequency channel of its power spectrogram is periodic of period $\frac{T_0}{H}$, where H is the hop size used for the STFT. Similarly, a *locally* spectrally-periodic signal b can be defined as a signal such that each frequency channel of its power spectrogram B is *locally* periodic, as follows:

$$\forall(t,f), \forall k \in [-K\cdots K], B(f,t) = B\left(f, t + k\frac{T_0(t)}{H}\right) \quad (2)$$

where $T_0(t)$ is the local spectral-period of the signal in seconds at time t and $K \in \mathbb{N}$ defines the range of time frames on which $T_0(t)$ can be approximated as constant.

Note that although we assumed that the voice does not exhibit an immediate repeating structure, it has nevertheless some periodicity, but generally at the pitch level ($\ll 1$ s) and the chorus level ($\gg 5$ s).

2.3. Derivation of the TF Mask

Let us assume that an estimate $\hat{B}$ of the power spectrogram of the background is available. We will focus on its estimation in section 3.2. Given X and $\hat{B}$ only, is it possible to derive a good TF mask W_b? Obviously, not having any particular model for V prevents a full rigorous probabilistic derivation of $W_b|\hat{B}$ and this problem is part of our current work. For now, we will hence focus on a *heuristic* method that proves to give very satisfying results in practice.

Conceptually, if $\hat{B}$ and X are very close for some TF bin (f,t), the energy in that bin is most likely to come from the background. On the contrary, if they are very different and in particular if $X(f,t) \gg \hat{B}(f,t)$, then the probability is high that the energy of this bin rather comes from the foreground signal (the voice). In [13], X and B are compared through $\rho(f,t) = \left|\log\frac{X(f,t)}{\hat{B}(f,t)}\right|$ and $W_b(f,t)$ is set to 1 if $\rho(f,t)$ stands below a given threshold called *tolerance*. Otherwise, $W_b(f,t)$ is set to 0, thus leading to a binary mask. The rationale underlying this choice of ρ is that the perception of sound is widely acknowledged to be related to log-spectral energy distribution.

In this study, we will concentrate on another expression for W_b based on a Gaussian radial basis function, that allows the mapping of ρ to the interval $[0\,1]$. This leads to a soft mask, which, unlike a binary mask, helps to reduce noise artifacts.

$$W_b(f,t) = \exp\left(-\frac{\left(\log X(f,t) - \log\hat{B}(f,t)\right)^2}{2\lambda^2}\right) \quad (3)$$

where λ is called the *tolerance* and is a parameter of the algorithm.

3. ESTIMATION

3.1. Time-Varying Repeating Period

In [13], the background signal was assumed to be only spectrally-periodic, i.e. with a fixed repeating period for all time frames. Here, we have assumed the background signal b to be locally spectrally-periodic, i.e. with a time-varying period $T_0(t)$. This generalization of REPET allows us to deal with long recordings, where the repeating background is likely to vary over time (e.g. verse vs. chorus).

To model the repeating background b, we first need to track its period $T_0(t)$. To do so, we compute the *beat spectrogram*, a two-dimensional representation of the sound that reveals the rhythmic variations over time, a concept originally introduced in [7]. Given the power spectrogram X of the mixture, we calculate a power spectrogram for each of its frequency channels. This gives the modulations of the energy for each of the frequency channels. The beat spectrogram Ω_X of the mixture is then defined as the average of the power spectrograms of all the frequency channels of X, as follows[1]:

$$\Omega_X = \frac{1}{F}\sum_{f=1}^{F}\left|\mathcal{F}_2\left(\overline{X}(f,.)\right)\right|^2 \quad (4)$$

where $\overline{X}(f,.)$ is the f^{th} frequency channel of X whose sliding mean has been removed and $\mathcal{F}_2$ is an STFT transform, with different parameters than $\mathcal{F}$ (see section 4.2 for the numerical values).

The computation of the beat spectrogram is depicted in Fig. 1.

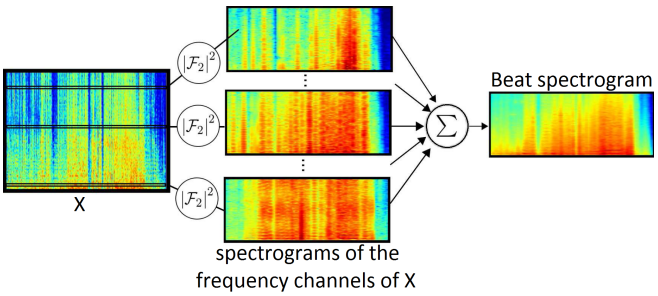

Fig. 1. Illustration of the computation of the beat spectrogram.

Given the beat spectrogram Ω_X, any method to estimate a time-varying prominent period can be used. Hence, we do not linger here on the details of the algorithm we used, but only outline its main

[1]Note that a further development of the method may include different spectral-periods for different frequency bands.

ideas[2]. The likelihood for each possible spectral-period and for each time slot was computed using a weighted spectral sum. The spectral-period is then obtained using a dynamic program that can be understood as a smoothing of these likelihoods. Values of $\{T_0(t)\}_{t=1\cdots T}$ are then obtained for all time frames through interpolation.

3.2. Repeating Background

We assume the background signal b is locally spectrally-periodic so that (2) holds. Furthermore, we assume its parameter K is known and its local spectral-period $T_0(t)$ has been computed for each time frame t as presented in section 3.1. Let $t_0(t) = \frac{T_0(t)}{H}$ where H is the hop size of the STFT operator $\mathcal{F}$.

In order to estimate B from X, we further assume that the lead signal is *sparse* in the TF domain, i.e. only a small portion of its TF representation contains values of a non-negligible magnitude, a reasonable assumption for voice signals. Hence, there are only a small amount of TF bins such that B strongly differs from X. Still, for the TF bins where the lead signal is active, the difference between B and X becomes significant. Thus, for a TF bin (f,t), it is likely that *most* $k \in [-K \cdots K]$ obey $B(f,t) \approx X(f, t+kt_0(t))$ while the others can be considered as *outliers*. from the perspective of estimating $B(f,t)$. For these reasons, robust estimation of $B(f,t)$ can be performed by computing the *median* value of $[X(f, t-Kt_0(t)) X(f, t-(K-1)t_0(t)) \cdots X(f, t+Kt_0(t))]$. The median is indeed known to be less sensitive to outliers.

A further refinement that proved to improve performance is to also assume that the background signal cannot have stronger energy than the mixture for any TF bin. This assumption comes from the fact that, given two independent processes B and V with zero means, we have $X \approx B + V$. Finally, the estimate $\hat{B}$ of B is given as:

$$B_0(f,t) = \text{median}\left[X(f, t-Kt_0(t)) \cdots X(f, t+Kt_0(t))\right]$$
$$\hat{B}(f,t) = \min\left(X(f,t), B_0(f,t)\right)$$

The TF mask W_b can then be computed using Eq. 3 and the separation can be performed using Eq. 1. The whole process only involves simple operations and can be very efficiently implemented.

4. EVALUATION

4.1. Dataset & Competitive Method

Recently, FitzGerald *et al* proposed the Multipass Median Filtering-based Separation (MMFS) method, a rather simple and novel approach for music/voice separation. Their approach is based on a median filtering of the 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 [6]. To evaluate their method, they fortunately found recordings released by the pop band *The Beach Boys*, where some of the complete original accompaniments and vocals were made available as split stereo tracks[3] and separated tracks[4]. After resynchronizing the accompaniments and vocals for the latter case, we created a total of 14 sources in the form of split stereo wave files sampled at 44.1 kHz, with the complete accompaniment and vocals on the left and right channels, respectively. As done in [6], we then used those 14 stereo sources to create three datasets of 14 mono mixtures, by mixing the channels at three different voice-to-music ratios: -6 dB (music is louder), 0 dB (original equivalent level), and 6 dB (voice is louder).

We decided to compare our extended version of REPET to the best version of the MMFS algorithm (there are 4 [6]), first because a dataset of complete real-world recordings was finally accessible for a comparative study, and then because we thought it could be interesting to compare two relatively simple and novel music/voice separation approaches. Note that although we are claiming to conduct a comparative study, we are not using the exact same dataset since first FitzGerald *et al* did not mention which tracks they used for their experiments, and also because unlike them, we chose to process the complete tracks without segmenting them, since our extended REPET can now handle longer audio recordings with varying repeating background, and this without computational constraints. Note also that we did not compare this extended version of REPET to the original one introduced in [13] since it does not make sense to apply the latter one on full tracks.

4.2. Parameters & Separation Measures

In the analysis stage, the STFT of each mixture was computed using a window length of 40 ms with 80% of overlapping. The beat spectrogram was computed using a window length of 10 seconds with 75% of overlapping. In the separation stage, each mixture was then processed by the REPET algorithm. The parameters λ and K were fixed to 1.5 and 2, respectively. In the masking stage, we used both a binary TF mask and the soft TF mask described in Eq. 3. As done in [6], we also applied a high-pass at 100 Hz on the vocal estimates.

We used the *BSS_EVAL toolbox* provided by [4] to measure the separation performance of our REPET algorithm. 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). As done in [6], we decided to measure SDR, SIR, and SAR on segments of 45 seconds from the music and voice estimates. Higher values of SDR, SIR, and SAR would imply better separation.

4.3. Results & Statistical Analysis

First, we compared the results of REPET with binary mask vs. soft mask, and without high-pass vs. with high-pass. A (non-parametric) Kruskal-Wallis one-way analysis of variance showed that using a high-pass at 100 Hz on the voice estimates gave overall statistically better results, except for the voice SAR. Furthermore, using a soft mask gave overall slightly better results, except for the voice SIR. The improvement was however statistically not significant, except for the voice SAR. We nevertheless believe that the estimates sound perceptually better when using a soft mask instead of a binary mask, therefore we decided to show the results only for the soft mask.

Since FitzGerald *et al* did not mention which tracks they used and only provided mean values, we could not conduct a statistical analysis to compare the results. We can however compare their means with our means and standard deviations, in the form of error bars. Thus, Fig. 2 and 3 show the average SDR, SIR, and SAR for the music and the voice estimates, respectively, at voice-to-music ratios of -6, 0, and 6 dB, without and with High-Pass at 100 Hz. The means and standard deviations of REPET are represented by the error bars and the means of MMFS are represented by the crosses.

In Fig. 2, we can see that for the music estimate, REPET has overall a lower SAR, but a higher SIR, and a similar SDR. This could mean that REPET is better for removing the vocal "interferences"

[2]The Python code for this algorithm is freely available under a GPL license at `http://www.telecom-paristech.fr/~liutkus`.

[3]Good Vibrations: Thirty Years of The Beach Boys, 1993

[4]The Pet Sounds Sessions, 1997

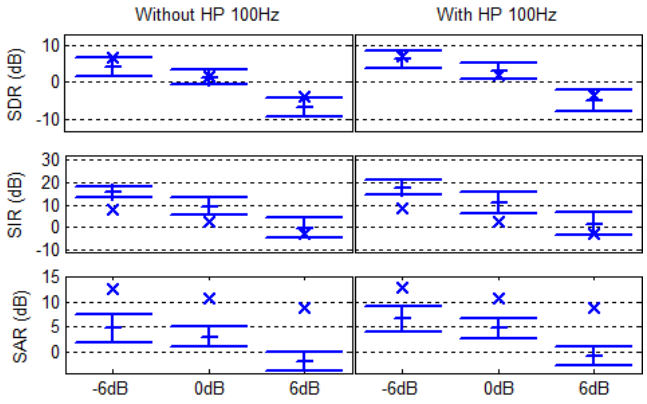

Fig. 2. SDR, SIR, and SAR of the *music* estimates, at voice-to-music mixing ratios of -6 dB, 0 dB, and 6 dB, without and with High-Pass at 100 Hz. The error bars represent the means (short horizontal lines in the middle) plus/minus standard deviations (long horizontal lines on each side) of REPET, while the crosses represent the means of the best MMFS. Higher values mean better separation.

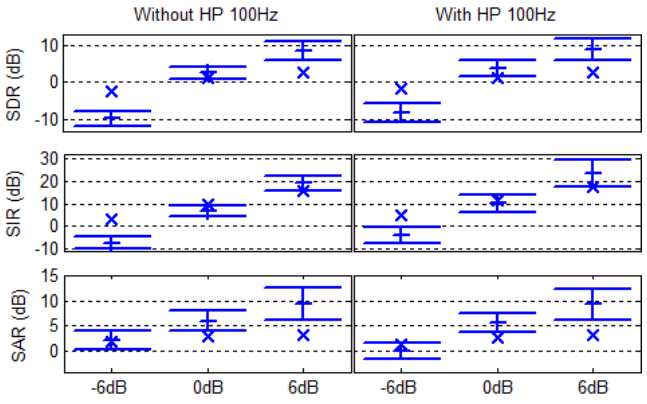

Fig. 3. SDR, SIR, and SAR of the *voice* estimates. (see Fig. 2)

from the accompaniment, at the price of introducing separation artifacts. In Fig. 3, we can see that for the voice estimate, REPET has overall worse results when the voice is softer, but better results when the voice is louder. This could mean that REPET is better at extracting the foreground outliers (vocals) from the repeating background (accompaniment) when there are larger in number.

The average computation time for our music/voice separation system over all the mixtures was 1.1830 s for 1 s of mixture, when implemented in *Python* on a PC with a dual-core processor and 8GB of RAM. As we can see, this extended REPET performs overall at least as well as a recent competitive music/voice separation method, but on complete recordings, while being computationally efficient.

5. CONCLUSION

In this study, we have presented an extension of the REPET algorithm for music/voice separation that allows processing of complete musical excerpts. The method is characterized by the assumption that the musical background exhibits *local* spectral-periodicity, which proved to be adequate for many kinds of musical tracks. Dropping absolute periodicity as was done in previous work permits to increase the expressive power of the model while remaining computationally tractable. Indeed, unlike other separation methods, REPET is only based on self-similarity. The algorithm is simple, fast, blind, and therefore completely and easily automatable.

Future work will include a more thorough probabilistic modeling and the ability to simultaneously separate several repeating patterns. Introducing dynamic features in source separation allows taking intuitive musicological knowledge into account and further refinements of the model may permit the user to manually specify the structure of the track to process in order to facilitate separation.

6. REFERENCES

[1] L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1):191–199, 2006.

[2] A.T. Cemgil, P. Peeling, O. Dikmen, and S. Godsill. Prior structures for Time-Frequency energy distributions. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 151–154, New Paltz, NY, USA, October 21-24 2007.

[3] J.-L. Durrieu, B. David, and G. Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1180–1191, October 2011.

[4] C. Févotte, R. Gribonval, and E. Vincent. BSS EVAL toolbox user guide. Technical Report 1706, IRISA, Rennes, France, April 2005. http://www.irisa.fr/metiss/bss eval/.

[5] D. FitzGerald, M. Cranitch, and E. Coyle. On the use of the beta divergence for musical source separation. In *16th IET Irish Signals and Systems Conference*, Galway, Ireland, June 18-19 2008.

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

[7] J. Foote and S. Uchihashi. The beat spectrum: A new approach to rhythm analysis. In *IEEE International Conference on Multimedia and Expo*, pages 881–884, Tokyo, Japan, August 22-25 2001.

[8] J. Han and C.-W. 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.

[9] C.-L. Hsu and J.-S. R. 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.

[10] Y. Li and D. 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.

[11] A. Liutkus, R. Badeau, and G. Richard. Gaussian processes for underdetermined source separation. *IEEE Transactions on Signal Processing*, 59(7):3155 –3167, 2011.

[12] A. Ozerov, P. Philippe, F. Bimbot, and R. 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.

[13] Z. Rafii and B. Pardo. A simple music/voice separation method 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.

[14] S. T. Roweis. One microphone source separation. In *Advances in Neural Information Processing Systems*, volume 13, pages 793–799. MIT Press, 2001.

[15] T. Virtanen, A. Mesaros, and M. 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, September 21 2008.