MUSIC-CONTENT-ADAPTIVE ROBUST PRINCIPAL COMPONENT ANALYSIS FOR A SEMANTICALLY CONSISTENT SEPARATION OF FOREGROUND AND BACKGROUND IN MUSIC AUDIO SIGNALS

Hélène Papadopoulos
Laboratoire des Signaux et Systèmes
UMR 8506, CNRS-SUPELEC-Univ. Paris-Sud, France
helene.papadopoulos[at]lss.supelec.fr

Daniel P.W. Ellis
LabROSA
Columbia University
dpwe[at]ee.columbia.edu

ABSTRACT

Robust Principal Component Analysis (RPCA) is a technique to decompose signals into sparse and low rank components, and has recently drawn the attention of the MIR field for the problem of separating leading vocals from accompaniment, with appealing results obtained on small excerpts of music. However, the performance of the method drops when processing entire music tracks. We present an adaptive formulation of RPCA that incorporates music content information to guide the decomposition. Experiments on a set of complete music tracks of various genres show that the proposed algorithm is able to better process entire pieces of music that may exhibit large variations in the music content, and compares favorably with the state-of-the-art.

1. INTRODUCTION

In the general context of processing high-dimensional data, a recurrent problem consists in extracting specific information from a massive amount of related or unrelated information. Examples include recovering documents with specific topics from a collection of Web text documents [1] or detecting moving objects from camera recordings for video surveillance purpose [2]. Among numerous existing methods, the technique of Robust Principal Component Analysis (RPCA) [3, 4], has recently drawn a lot of attention. All the above-mentioned problems can be formulated as separating some foreground components (the keywords in Web data, the moving objects in video) from an underlying background (the background corpus topic in Web data, the stable environment in video), that can be respectively modeled as a sparse plus a low-rank contribution.

RPCA has been used extensively in the field of image processing (e.g. image segmentation [5], visual pattern correspondence [6], surveillance video processing [7], batch image alignment [8], etc.). However, its application in Music Information Retrieval (MIR) is much more recent. Existing applications in audio include audio classification, as in [9] where audio segments from video sound files are classified into classes (applause and laughter occurrences); [10] addresses the problem of refining available social tags obtained through social tagging websites to maximize their quality. The main application of the RPCA framework in music focuses on the task of separating a foreground component, usually the singing voice, from a background accompaniment in monaural polyphonic recordings, i.e., when only one channel of recording is available. This scenario is the primary focus of this paper.

The singing voice is a complex and important music signal attribute that has been much studied in MIR. Its separation is essential for many applications, such as singer identification [11], melody transcription [12], or query by humming [13]. We refer the reader to [14] for a recent review of singing voice separation methods. Recently, approaches that take advantage of repetition in the signal have emerged. These approaches assume that the background accompaniment has a repetitive musical structure, in contrast to the vocal signal whose repetitions, if any, occur only at a much larger timescale [15, 16, 17]. In [15] a simple method for separating music and voice is proposed based on the extraction of the underlying repeating musical structure using binary time-frequency masking (REPET algorithm). The methods assumes that there is no variations in the background and is thus limited to short excerpts. In [16], the method is generalized to permit the processing of complete musical tracks by relying on the assumption of local spectral-periodicity. Moreover, artifacts are reduced by using soft-masks. Inspired by these approaches, [17] proposes a model for singing voice separation based on repetition, but without using the hypothesis of local periodicity. The background musical accompaniment at a given frame is identified using the nearest neighbor frames in the whole mixture spectrogram.

Most recently, RPCA has emerged as a promising approach to singing voice separation based on the idea that the repetitive musical accompaniment may lie in a low-rank subspace, while the singing voice is relatively sparse in the time-frequency domain [18]. The voice and the accompaniment are separated by decomposing the Short-Time-Fourier Transform (STFT) magnitude (i.e., spectrogram) into sparse and low-rank components. When tested on short audio excerpts from the MIR-1K dataset[1] RPCA shows improvement over two state-of-the-art approaches [19, 15]. The decomposition is improved in [20] by adding a regularization term to incorporate a prior tendency towards harmonicity in the low-rank component, reflecting the fact that background voices can be described as a harmonic series of sinusoids at multiples of a fundamental frequency. A post-processing step is applied to the sparse component of the decomposition to eliminate the percussive sounds. [21] addresses the problem of jointly finding a sparse approximation of a varying component (e.g., the singing voice) and a repeating background (e.g., the musical accompaniment) in the same *redundant dictionary*. In parallel with the RPCA idea of [3], the mixture is decomposed into a sum of two components: a *structured* sparse matrix and an *unstructured* sparse matrix. Structured sparsity is enforced using mixed norms, along

[1]The MIR-1K dataset [19] is a set of 1000 short excerpts ($4 - 13$s) extracted from 110 Chinese karaoke pop songs, where accompaniment and the singing voices are separately recorded. See https://sites.google.com/site/unvoicedsoundseparation/mir-1k.

with a greedy Matching Pursuit algorithm [22]. The model is evaluated on short popular music excerpts from the Beach Boys. [23] proposes a non-negative variant of RPCA, termed robust low-rank non-negative matrix factorization (RNMF). In this approach the low-rank model is represented as a non-negative linear combination of non-negative basis vectors. The proposed framework allows incorporating unsupervised, semi-, and fully-supervised learning, with supervised training drastically improving the results of the separation. Other related works including [24, 25] address singing voice separation based on low-rank representations alone but are beyond the scope of this article.

While RPCA performs well on the ∼10 sec clips of MIR-1K, the full-length Beach Boys examples of [14] give much less satisfying results. When dealing with whole recordings, the musical background may include significant changes in instrumentation and dynamics which may rival the variation in the foreground, and hence its rank in the spectrogram representation. Further, foreground may vary in its complexity (e.g., solo voice followed by a duet) and may be unevenly distributed throughout the piece (e.g., entire segments with background only). Thus, the best way to apply RPCA to separate *complete* music pieces remains an open question.

In this article, we explore an adaptive version of RPCA (A-RPCA) that is able to handle complex music signals by taking into account the intrinsic musical content. We aim to adjust the task through the incorporation of domain knowledge that guides the decomposition towards results that are physically and musically meaningful. Time-frequency representations of music audio may be structured in several ways according to their content. For instance, the frequency axis can be segmented into regions corresponding to the spectral range of each instrument of the mixture. In the singing separation scenario, coefficients that are not in the singing voice spectral band should not be selected in the sparse layer. In the time dimension, music audio signals can generally be organized into a hierarchy of segments at different scales, each with its own semantic function (bar, phrase, entire section etc.), and each having specific characteristics in terms of instrumentation, leading voice, etc. Importantly, as the segments become shorter, we expect the accompaniment to span less variation, and thus the rank of the background to reduce.

We will show a way for this music content information to be incorporated in the decomposition to allow an accurate processing of *entire* music tracks. More specifically, we incorporate voice activity information as a cue to separate the leading voice from the background. Music pieces can be segmented into vocal segments (where the leading voice is present) and background segments (that can be purely instrumental or may contain backing voices). Finding vocal segments (voicing detection [26]) is a subject that has received significant attention within MIR [26, 27, 28, 29]. The decomposition into sparse and low-rank components should be coherent with the semantic structure of the piece: the sparse (foreground) component should be denser in sections containing the leading voice while portions of the sparse matrix corresponding to non-singing segments should ideally be null. Thus, while the technique remains the same as [18] at the lowest level, we consider the problem of segmenting a longer track into suitable pieces, and how to locally adapt the parameters of the decomposition by incorporating prior information.

2. ROBUST PRINCIPAL COMPONENT ANALYSIS VIA PRINCIPAL COMPONENT PURSUIT

In [3] , Candès *et al.* show that, under very broad conditions, a data matrix $D \in \mathbb{R}^{m\times n}$ can be exactly and uniquely decomposed into a low-rank component A and a sparse component E via a convex program called *Principal Component Pursuit* (RPCA-PCP) given by:

$$\min_{A,E} \quad \|A\|_* + \lambda\|E\|_1 \quad \text{s.t.} \quad D = A + E \tag{1}$$

where $\lambda > 0$ is a regularization parameter that trades between the rank of A and the sparsity of E. The nuclear norm $\|\cdot\|_*$ – the sum of singular values – is used as surrogate for the rank of A [30], and the ℓ_1 norm $\|\cdot\|_1$ (sum of absolute values of the matrix entries) is an effective surrogate for the ℓ_0 pseudo-norm, the number of non-zero entries in the matrix [31, 32].

The Augmented Lagrange Multiplier Method (ALM) and its practical variant, the Alternating Direction Method of Multipliers (ADM), have been proposed as efficient optimization schemes to solve this problem [33, 34, 35]. ALM works by minimizing the augmented Lagrangian function of (1):

$$\mathcal{L}(A, E, Y, \mu) = \|A\|_* + \lambda\|E\|_1 + \langle Y, A+E-D\rangle + \frac{\mu}{2}\|A+E-D\|_F^2 \tag{2}$$

where $Y \in \mathbb{R}^{m\times n}$ is the Lagrange multiplier of the linear constraint that allows removing the equality constraint, $\mu > 0$ is a penalty parameter for the violation of the linear constraint, $\langle\cdot,\cdot\rangle$ denotes the standard trace inner product and $\|\cdot\|_F$ is the Frobenius norm[2]. ALM [34] is an iterative scheme that works by repeatedly minimizing A and E simultaneously. In contrast, ADM splits the minimization of (2) into two smaller and easier subproblems, with A and E minimized sequentially:

$$A^{k+1} = \operatorname*{argmin}_A \quad \mathcal{L}(A, E^k, Y^k, \mu^k) \tag{3a}$$

$$E^{k+1} = \operatorname*{argmin}_E \quad \mathcal{L}(A^{k+1}, E, Y^k, \mu^k) \tag{3b}$$

Both subproblems (3a) and (3b) are shrinkage problems that have closed-form solutions that we briefly present here. We refer the reader to [34, 35] for more details. For convenience we introduce the scalar soft-thresholding (shrinkage) operator $\mathcal{S}_\epsilon[x]$:

$$\mathcal{S}_\epsilon[x] = \text{sgn}(x)\cdot\max(|x|-\epsilon, 0) = \begin{cases} x-\epsilon & \text{if} \quad x > \epsilon \\ x+\epsilon & \text{if} \quad x < -\epsilon \\ 0 & \text{otherwise} \end{cases}$$

where $x \in \mathbb{R}$ and $\epsilon > 0$. This operator can be extended to matrices by applying it element-wise.

Problem (3a) is equivalent to:

$$A^{k+1} = \min_A \left\{ \|A\|_* + \frac{\mu^k}{2}\|A - (D - E^k + \frac{1}{\mu^k}Y^k)\|_F^2 \right\} \tag{4}$$

that has, according to [36], a closed-from solution given by:

$$A^{k+1} = U\mathcal{S}_{\frac{1}{\mu^k}}[\Sigma]V^T$$

[2]The Frobenius norm of matrix A is defined as $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$.

where $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$ and $\Sigma \in \mathbb{R}^{r\times r}$ are obtained via the singular value decomposition $(U, \Sigma, V) = SVD(D - E^k + \frac{Y^k}{\mu^k})$.

Problem (3b) can be written as:

$$E^{k+1} = \min_E \left\{ \lambda \|E\|_1 + \frac{\mu^k}{2} \|E - (D - A^{k+1} + \frac{1}{\mu^k} Y^k)\|_F^2 \right\} \quad (5)$$

whose solution is given by the least-absolute shrinkage and selection operator (*Lasso*) [37], a method also known in the signal processing community as basis pursuit denoising [38]:

$$E^{k+1} = \mathcal{S}_{\frac{\lambda}{\mu_k}}[D - A^{k+1} + \frac{Y^k}{\mu_k}]$$

In other words, denoting $G^E = D - A^{k+1} + \frac{Y^k}{\mu^k}$:

$$\forall\, i \in [1, m], \forall\, j \in [1, n] \quad E_{ij}^{k+1} = \text{sgn}(G_{ij}^E) \cdot \max(|G_{ij}^E| - \frac{\lambda}{\mu^k}, 0)$$

3. ADAPTIVE RPCA (A-RPCA)

As discussed in Section 1, in a given song, the foreground vocals typically exhibit a clustered distribution in the time-frequency plane relating to the semantic structure of the piece that alternates between vocal and non-vocal (background) segments. This structure should be reflected in the decomposition: frames belonging to singing voice-inactive segments should result in zero-valued columns in E.

The balance between the sparse and low-rank contributions is set by the value of the regularization parameter λ. The voice separation quality with respect to the value of λ for the *Pink Noise Party* song *Their Shallow Singularity* is illustrated in Fig. 1. As we can observe, the best λ differs depending on whether we process the entire song, or restrict processing to just the singing voice-active parts. Because the separation for the background part is monotonically better as λ increases, the difference between the optimum λ indicates that the global separation quality is compromised between the singing voice and the background part.

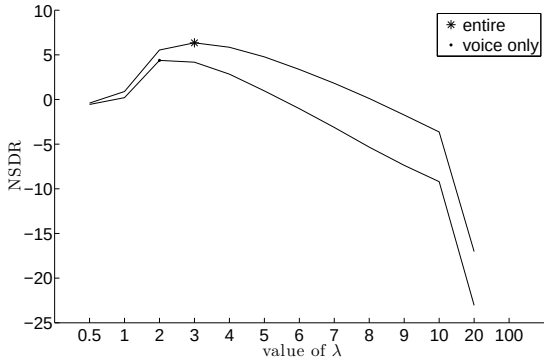

Figure 1: Variation of the estimated singing voice NSDR (see definition in Section 4) according to the value of λ under two situations. •: NSDR when only the singing voice-active parts of the separated signal are processed. $*$: NSDR when the entire signal is processed.

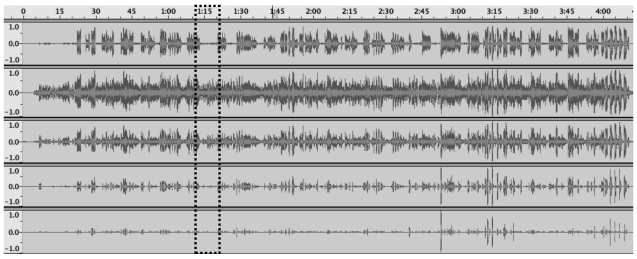

Figure 2: Waveform of the separated voice for various values of λ for the song *Is This Love* by Bob Marley. From top to bottom: clean voice, $\lambda = \lambda_1, 2 * \lambda_1, 5 * \lambda_1, 10 * \lambda_1$.

In the theoretical formulation of RPCA-PCP [3], there is no single value of λ that works for separating sparse from low-rank components in all conditions. They recommend $\lambda = \max(m, n)^{-\frac{1}{2}}$ but also note that the decomposition can be improved by choosing λ in light of prior knowledge about the solution. In practice, we have found that the decomposition of music audio is very sensitive to the choice of λ with frequently no single value able to achieve a satisfying separation between voice and instrumental parts across a whole recording. This is illustrated in Fig. 2, which shows the waveforms of the resynthesized separated voice obtained with the RPCA-PCP formulation for various λ. For $\lambda = \lambda_1 = 1/\sqrt{\max(m, n)}$ and $\lambda_2 = 2{*}\lambda_1$, around $t = 1.15$ s (dashed rectangle) there is a non-zero contribution in the voice layer but no actual lead vocal. This is eliminated with $\lambda = 5 * \lambda_1, 10 * \lambda_1$ but at the expense of a very poor quality voice estimate: the resulting signal consists of percussive sounds and higher harmonics of the instruments, and does not resemble the voice. Note that similar observations have been made in the context of video surveillance [39].

To address the problem of variations in λ, we propose an adaptive variant of the RPCA consisting of a weighted decomposition that incorporates prior information about the music content. Specifically, voice activity information is used as a cue to adjust the regularization parameter through the entire analyzed piece in the (3b) step, and therefore better match the balance between sparse and low-rank contributions to suit to the actual music content. This idea is related to previous theoretical work [40, 41, 42], but to our knowledge, its application in the framework of RPCA is new.

We consider a time segmentation of the magnitude spectrogram into N_{block} consecutive (non-overlapping) blocks of vocal / non-vocal (background accompaniment) segments. We can represent the magnitude spectrogram as a concatenation of column-blocks $D = [D_1 D_2 \cdots D_{N_{\text{block}}}]$, the sparse layer as $E = [E_1 \cdots E_{N_{\text{block}}}]$ and $G^E = [G_1^E \cdots G_{N_{\text{block}}}^E]$.

We can minimize the objective function with respect to each column-block separately. To guide the separation, we aim at setting a different value of $\lambda_l, l \in [1, N_{blocks}]$ for each block according to the voice activity side information. For each block, the problem is equivalent to Eq. (5) and accordingly, the solution to the resulting problem:

$$E_l^{k+1} = \min_{E_l} \left\{ \lambda_l \|E_l\|_1 + \frac{\mu^k}{2} \|E_l - G_l^E\|_F^2 \right\}$$

is given by:

$$E_l^{k+1} = \mathcal{S}_{\frac{\lambda_l}{\mu_k}}[G_l^E] \quad (6)$$

Algorithm 1 Adaptive RPCA (A-RPCA)

Input: spectrogram D, blocks, $\lambda, \lambda_1, \ldots, \lambda_{N_{blocks}}$
Output: E, A
Initialization: $Y^0 = D/J(D)$ where $J(D) = max(\|D\|_2, \lambda^{-1}\|D\|_\infty)$; $E^0 = 0$; $\mu_0 > 0$; $\rho > 1$; $k = 0$
while not converged **do**
  update A:
  $(U, \Sigma, V) = SVD(D - E^k + \frac{Y^k}{\mu^k})$; $A^{k+1} = U\mathcal{S}_{\frac{1}{\mu^k}}[\Sigma]V^T$
  update E:
  for each block l **do**
    $\lambda = \lambda_l$;
    $E_l^{t+1} = \mathcal{S}_{\frac{\lambda_l}{\mu_k}}[D_l - A_l^{k+1} + \frac{Y_l^k}{\mu^k}]$
  end for
  $E^{t+1} = [E_1^{t+1} E_2^{t+1} \cdots E_{N_{\text{block}}}^{t+1}]$
  update Y, μ:
  $Y^{k+1} = Y^k - \mu^k(A^{k+1} + E^{k+1} - D)$
  $\mu^{k+1} = \rho \cdot \mu^k$
  $k = k + 1$
end while

Denote λ_{v} the constant value of the regularization parameter λ used in the basic formulation of RPCA for voice separation [18]. To guide the separation, in the A-RPCA formulation we assign to each block a value λ_l in accordance with the considered prior music structure information. Using a large λ_l in blocks without leading voice will favor retaining non-zero coefficients in the accompaniment layer. Denoting by Ω_V the set of time frames that contain voice, the values of λ_l are set as:

$$\forall\, l \in [1, N_{\text{block}}] \begin{cases} \lambda_l = \lambda_{\text{v}} \text{ if } E_l \subset \Omega_V \\ \lambda_l = \lambda_{\text{nv}} \text{ otherwise} \end{cases} \quad (7)$$

with $\lambda_{\text{nv}} > \lambda_{\text{v}}$ to enhance sparsity of E when no vocal activity is detected. Note that instead of two distinct values of λ_l, further improvements could be obtained by tuning λ_l more precisely to suit the segment characteristics. For instance, vibrato information could be used to quantify the amount of voice in the mixture within each block and to set a specific regularization parameter accordingly. The update rules of the A-RPCA algorithm are detailed in Algorithm 1.

In Section 4, we investigate the results of adaptive-RPCA with both exact (ground-truth) and estimated vocal activity information. For estimating vocal activity information, we use the voicing detection step of the melody extraction algorithm implemented in the MELODIA Melody Extraction vamp plug-in[3], as it is freely available for people to download and use. We refer the reader to [26] and references therein for other voicing detection algorithms. The algorithm for the automatic extraction of the main melody from polyphonic music recordings implemented in MELODIA is a salience-based model that is described in [43]. It is based on the creation and characterization of pitch contours grouped using auditory streaming cues, and includes a voice detection step that indicates when the melody is present; we use this melody location as an indicator of leading voice activity. Note that while melody can sometimes be carried by other instruments, in the evaluation dataset of Section 4 it is mainly singing.

[3] `http://mtg.upf.edu/technologies/melodia`

4. EVALUATION

In this section, we present the results of our approach evaluated on a database of complete music tracks of various genres. We compare the proposed adaptive method with the baseline method [18] as well as another state-of-the-art method [16]. Sound examples discussed in the article can be found at:
`http://papadopoulosellisdafx14.blogspot.fr`.

4.1. Parameters, Dataset and Evaluation Criteria

To evaluate the proposed approach, we have constructed a database of 12 complete music tracks of various genres, with separated vocal and accompaniment files, as well as mixture versions formed as the sum of the vocal and accompaniment files. The tracks, listed in Tab. 1, were created from multitracks mixed in Audacity[4], then exported with or without the vocal or accompaniment lines.

Following previous work [18, 44, 15], the separations are evaluated with metrics from the BSS-EVAL toolbox [45], which provides a framework for the evaluation of source separation algorithms when the original sources are available for comparison. Three ratios are considered for both sources: Source-to-Distortion (SDR), Sources-to-Interference (SIR), and Sources-to-Artifacts (SAR). In addition, we measure the improvement in SDR between the mixture d and the estimated resynthesized singing voice $\hat{e}$ by the Normalized SDR (NSDR, also known as *SDR improvement*, SDRI), defined for the voice as $\text{NSDR}(\hat{e}, e, d) = \text{SDR}(\hat{e}, e) - \text{SDR}(d, e)$, where e is the original clean singing voice. The same measure is used for the evaluation of the background. Each measure is computed globally on the whole track, but also locally according to the segmentation into vocal/non-vocal segments. Higher values of the metrics indicate better separation.

We compare the results of the A-RPCA with musically-informed adaptive λ and the baseline RPCA method [18] with fixed λ, using the same parameter settings in the analysis stage: the STFT of each mixture is computed using a window length of 1024 samples with 75% overlap at a sampling rate of 11.5KHz. No post-processing (such as masking) is added. After spectrogram decomposition, the signals are reconstructed using the inverse STFT and the phase of the original signal.

The parameter λ is set to $1/\sqrt{\max(m, n)}$ in the baseline method. Two different versions of the proposed A-RPCA algorithm are evaluated. First, A-RPCA with exact voice activity information, using manually annotated ground-truth (A-RPCA_GT), and $\lambda_l = \lambda$ for singing voice regions and $\lambda_l = 5 * \lambda$ for background only regions. In the other configuration, estimated voice activity location is used (A-RPCA_est), with same settings for the λ_l.

We also compare our approach with the REPET state-of-the-art algorithm based on repeating pattern discovery and binary time-frequency masking [16]. Note that we use for comparison the version of REPET that is designed for processing complete musical tracks (as opposed to the original one introduced in [15]). This method includes a simple low pass filtering post-processing step [46] that consists in removing all frequencies below 100Hz from the vocal signal and adding these components back into the background layer. We further apply this post-processing step to our model before comparison with the REPET algorithm.

Paired sample t-tests at the 5% significance level are performed to determine whether there is statistical significance in the results between various configurations.

[4] http://audacity.sourceforge.net

Table 1: Sound excerpts used for the evaluation; *back.* proportion of background (no leading voice) segments (in % of the whole excerpt duration); Recall *Rec.* and False Alarm *F.A.* voicing detection rate.

Name	% *back.*	*Rec.*	*F.A.*	Name	% *back.*	*Rec.*	*F.A.*
1- *Beatles* Sgt Pepper's Lonely Hearts Club Band	49.3	74.74	45.56	8 - *Bob Marley* Is This Love	37.2	66.22	36.84
2 - *Beatles* With A Little Help From My Friends	13.5	70.10	14.71	9 - *Doobie Brothers* Long Train Running	65.6	84.12	58.51
3 - *Beatles* She's Leaving Home	24.6	77.52	30.17	10 - *Marvin Gaye* Heard it Through The Grapevine	30.2	79.22	17.90
4 - *Beatles* A Day in The Life	35.6	61.30	63.96	11 - *The Eagles* Take it Easy	35.5	78.68	30.20
5,6 - *Puccini* piece for soprano and piano	24.7	47.90	27.04	12 - *The Police* Message in aBottle	24.9	73.90	20.44
7 - *Pink Noise Party* Their Shallow Singularity	42.1	64.15	61.83				

4.2. Results and Discussion

Results of the separation for the sparse (singing voice) and low-rank (background accompaniment) layers are presented in Tables 2, 3, 4 and 5. To have a better insight of the results we present measures computed both on the entire song and on the singing voice-active part only, that is obtained by concatenating all segments labeled as vocal segments in the ground truth.

• **Global separation results.** As we can see from Tables 2 and 3, using a musically-informed adaptive regularization parameter allows improving the results of the separation both for the background and the leading voice components. Note that the larger the proportion of purely-instrumental segments in a piece (see Tab. 1), the larger the results improvement (see in particular pieces 1, 7, 8 and 9), which is consistent with the goal of the proposed method. Statistical tests show that the improvement in the results is significant.

As discussed in Section 3, the quality of the separation with the baseline method [18] depends on the value of the regularization parameter. Moreover, the value that leads to the best separation quality differs from one music excerpt to another. Thus, when processing automatically a collection of music tracks, the choice of this value results from a trade-off. We report here results obtained with the typical choice $\lambda_{\text{v}} = 1/\sqrt{\max(m,n)}$ in Eq. (7). Note that for a given value of λ_{v} in the baseline method, the separation can always be further improved by the A-RPCA algorithm using a regularization parameter that is adapted to the music content based on prior music structure information: in all experiments, for a given constant value λ_{v} in the baseline method, setting $\lambda_{\text{nv}} > \lambda_{\text{v}}$ in Eq. (7) improves the results.

For the singing voice layer, improved SDR (better overall separation performance) and SIR (better capability of removing music interferences from the singing voice) with A-RPCA are obtained at the price of introducing more artifacts in the estimated voice (lower SAR_{voice}). Listening tests reveal that in some segments processed by A-RPCA, as for instance segment $[1'00'' - 1'15'']$ in Fig. 3, one can hear some high frequency isolated coefficients superimposed to the separated voice. This drawback could be reduced by including harmonicity priors in the sparse component of RPCA, as proposed in [20]. This performance trade-off is commonly encountered in music/voice separation [14, 47]. However, we can notice that all three measures are significantly improved with A-RPCA for the background layer.

• **Ground truth versus estimated voice activity location.** Imperfect voice activity location information still allows an improvement, although to a lesser extent than with ground-truth voice activity information. In table 1, we report the accuracy results of the voicing detection step. Similarly to the measures used for melody detection in [48, 12], we consider the *Voicing Recall Rate*, defined as the proportion of frames labeled voiced in the ground truth that are estimated as voiced frames by the algorithm, and the *Voicing False Alarm Rate*, defined as the proportion of frames labeled as unvoiced in the ground truth that are mistakenly estimated to be voiced by the algorithm. The decrease in the results mainly comes from background segments classified as vocal segments. However, statistical tests show that the improvement in the results between RPCA and A-RPCA_est is still significant.

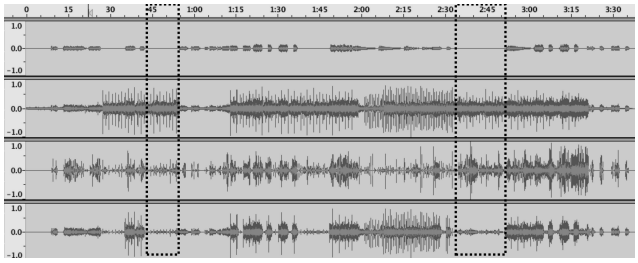

Figure 3: Separated voice for various values of λ for the *Pink Noise Party* song *Their Shallow Singularity*. From top to bottom: clean voice, constant $\lambda_1 = 1/\sqrt{\max(m,n)}$, constant $\lambda = 5 * \lambda_1$, adaptive $\lambda = (\lambda_1, 5 * \lambda_1)$.

• **Local separation results.** It is interesting to note that using an adaptive regularization parameter in a unified analysis of the whole piece is different from separately analyzing the successive vocal/non-vocal segments with different but constant values of λ (see for instance the dashed rectangles areas in Fig. 3).

• **Analysis of the results on vocal segments:** We expect the separation on background-only parts of the song to be improved with the A-RPCA algorithm. Indeed the side information directly indicates these regions where the foreground (sparse) components should be avoided; this can be clearly seen in Fig. 3. However, the improvements under the proposed model are not limited to non-vocal regions only. Results measured on the vocal segments alone indicate that by using the adaptive algorithm, the voice is also better estimated, as shown in Table 3. The improvement over RPCA is statistically significant, both when using ground truth and estimated voice activity location information. This indicates that side information helps not only to better determine the background only segments, but also enables improved recovery of the singing voice, presumably because the low-rank background model is a better match to the actual background.

Side information could have be added as a pre- or post-processing step to the RPCA algorithm. The adaptive-RPCA algorithm presents advantages over such approaches. To analyze this, we compare the

Table 2: SDR, SIR and SAR (in dB) and NSDR results for the voice (*Voice*) and background layer (*Back.*), computed across the whole song, for all models, averaged across all the songs. RPCA is the baseline system, A-RPCA_GT is the adaptive version using ground truth voice activity information, and A-RPCA_est uses estimated voice activity.

Entire song				
		RPCA	A-RPCA_GT	A-RPCA_est
Voice	SDR (dB)	-4.66	-2.16	-3.18
	SIR (dB)	-3.86	0.74	-0.46
	SAR (dB)	8.99	4.81	3.94
	NSDR	1.70	4.20	3.18
Back.	SDR (dB)	4.14	6.52	6.08
	SIR (dB)	11.48	13.30	12.07
	SAR (dB)	5.51	8.03	7.83
	NSDR	-2.35	0.03	-0.41

Table 3: SDR, SIR and SAR (in dB) and NSDR results for the voice (*Voice*) and background layer (*Back.*), computed across the vocal segments only, for all models, averaged across all the songs. RPCA is the baseline system, A-RPCA_GT is the adaptive version using ground truth voice activity information, and A-RPCA_est uses estimated voice activity.

Vocal segments				
		RPCA	A-RPCA_GT	A-RPCA_est
Voice	SDR (dB)	-3.19	-2.00	-1.96
	SIR (dB)	-2.33	-0.39	0.74
	SAR (dB)	9.44	7.27	4.64
	NSDR	1.67	2.85	2.90
Back.	SDR (dB)	3.63	5.18	5.28
	SIR (dB)	9.95	10.64	10.41
	SAR (dB)	5.39	7.32	7.54
	NSDR	-1.37	0.18	0.29

Table 4: SDR, SIR and SAR (in dB) and NSDR results for the voice (*Voice*) and background layer (*Back.*), computed across the whole song, for all models, averaged across all the songs. RPCA is the baseline system, A-RPCA_GT is the adaptive version using ground truth voice activity information, and A-RPCA_est uses estimated voice activity. Low-pass filtering post-processing is applied. REPET is the comparison algorithm [16].

Entire song					
		RPCA	A-RPCA_GT	A-RPCA_est	REPET
Voice	SDR (dB)	-2.76	-0.72	-2.11	-2.20
	SIR (dB)	-0.17	4.03	2.22	1.34
	SAR (dB)	4.33	3.33	2.32	3.19
	NSDR	3.60	5.64	4.25	4.16
Back.	SDR (dB)	5.16	7.61	6.81	5.01
	SIR (dB)	14.53	14.49	12.99	16.83
	SAR (dB)	5.96	9.02	8.44	5.47
	NSDR	-1.32	1.12	0.33	-1.48

Table 5: SDR, SIR and SAR (in dB) and NSDR results for the voice (*Voice*) and background layer (*Back.*), computed across the vocal segments only, for all models, averaged across all the songs. RPCA is the baseline system, A-RPCA_GT is the adaptive version using ground truth voice activity information, and A-RPCA_est uses estimated voice activity. Low-pass filtering post-processing is applied. REPET is the comparison algorithm [16].

Vocal segments only					
		RPCA	A-RPCA_GT	A-RPCA_est	REPET
Voice	SDR (dB)	-1.25	-0.53	-0.83	-0.70
	SIR (dB)	1.49	3.04	3.62	3.02
	SAR (dB)	5.02	4.46	3.12	4.02
	NSDR	3.60	4.32	4.02	4.15
Back.	SDR (dB)	4.85	6.03	6.11	4.80
	SIR (dB)	13.07	12.38	11.41	15.33
	SAR (dB)	5.91	7.69	8.20	5.41
	NSDR	-0.14	1.03	1.11	-0.20

A-RPCA algorithm with two variants of RPCA incorporating side information either as a pre- or a post-processing step:

- RPCA_OV_{pre}: Only the concatenation of segments classified as vocal is processed by RPCA (the singing voice estimate being set to zero in the remaining non-vocal segments).
- RPCA_OV_{post}: The whole song is processed by RPCA and non-zeros coefficients estimated as belonging to the voice layer in non-vocal segments are transferred to the background layer.

Results of the decomposition computed across the vocal segments only are presented in Table 6. Note that the RPCA_OV_{post} results reduce to the RPCA results in Table 3 since they are computed on vocal segments only. There is no statistical difference between the estimated voice obtained by processing with RPCA the whole song and the vocal segments only. Results are significantly better using the A-RPCA algorithm than using RPCA_OV_{pre} and RPCA_OV_{post}. This is illustrated in Figure 4, which shows an example of the decomposition on an excerpt of the *Doobie Brothers* song *Long Train Running* composed by a non-vocal followed by a vocal segment. We can see that there are misclassified partials in the voice spectrogram obtained with the baseline RPCA that are removed with A-RPCA. Moreover, the gap in the singing voice around frame 50 (breathing) is cleaner in the case of A-RPCA than in the case of RPCA. Listening tests confirm that the background is better attenuated in the voice layer when using A-RPCA.

Table 6: SDR, SIR and SAR (in dB) and NSDR results for the voice (*Voice*) and background layer (*Back.*), computed across the vocal segments only, averaged across all the songs. RPCA_OV_{post} is when using the baseline system and set the voice estimate to zero in background-only segments, RPCA_OV_{pre} is when processing only the voice segments with the baseline model, A-RPCA_GT is the adaptive version using ground truth voice activity information, and A-RPCA_est uses estimated voice activity.

		RPCA_OV_{post}	RPCA_OV_{pre}	A-RPCA_GT	A-RPCA_est
Voice	SDR	-3.19	-3.28	-2.00	-1.96
	SIR	-2.33	-2.31	3.62	0.74
	SAR	9.44	8.97	7.27	4.64
	NSDR	1.67	1.57	2.85	2.90
Back.	SDR	3.63	3.72	5.18	5.28
	SIR	9.95	9.22	10.64	10.41
	SAR	5.39	5.85	7.32	7.54
	NSDR	-1.37	-1.28	0.18	0.29

• **Comparison with the state-of-the-art.** As we can see from Table 4, the results obtained with the RPCA baseline method are not better than those obtained with the REPET algorithm. On the contrary, the REPET algorithm is significantly outperformed by the A-RPCA algorithm when using ground truth voice activity information, both for the sparse and low-rank layers. However, note that when using estimated voice activity information, the differ-

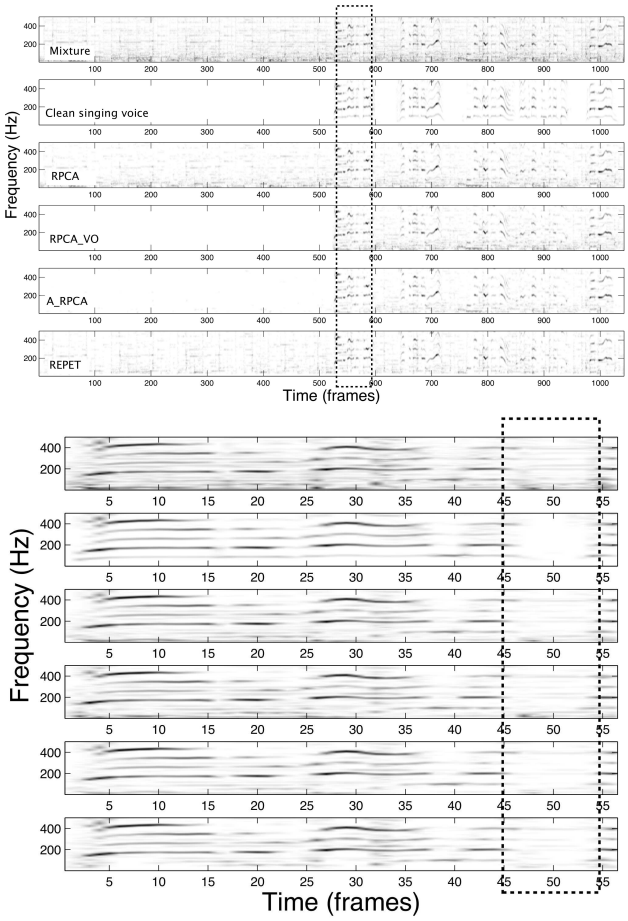

Figure 4: [Top Figure] Example decomposition on an excerpt of the *Doobie Brothers* song *Long Train Running* and [Bottom Figure] zoom between frames [525-580] (dashed rectangle in the Top Figure). For each figure, the top pane shows the part between 0 and 500Hz of the spectrogram of the original signal. The clean signing voice appears in the second pane. The separated signing voice obtained with baseline model (RPCA), with the baseline model when restricting the analysis to singing voice-active segments only (RPCA_OV_{pre}), and with the proposed A-RPCA model are represented in panes 3 to 5. For comparison, the sixth pane shows the results obtained with REPET [16].

ence in the results between REPET and A-RPCA is not statistically significant for the sparse layer. If we look closer at the results, it is interesting to note that the voice estimation improvement by A-RPCA_GT over REPET mainly comes from the non-vocal parts where the voice estimated is favored to be null. Indeed, Table 5 indicate that the voice estimates on vocal segments obtained with A-RPCA_GT and REPET are similar. This is illustrated by the two last panes in the [bottom] Figure 4, which show similar spectrograms of the voice estimates obtained with the A-RPCA and REPET algorithms on the vocal part of the excerpt.

5. CONCLUSION

We have explored an adaptive version of the RPCA technique that allows the processing of entire pieces of music including local variations in the music structure. Music content information is incorporated in the decomposition to guide the selection of coefficients in the sparse and low-rank layers according to the semantic structure of the piece. This motivates the choice of using a regularization parameter that is informed by musical cues. Results indicate that with the proposed algorithm, not only the background segments are better discriminated, but also that the singing voice is better estimated in vocal segments, presumably because the low-rank background model is a better match to the actual background. The method could be extended with other criteria (singer identification, vibrato saliency. etc.). It could also be improved by incorporating additional information to set differently the regularization parameters for *each* track to better accommodate the varying contrast of foreground and background. The idea of an adaptive decomposition could also be improved with a more complex formulation of RPCA that incorporates additional constraints [20] or a learned dictionary [49].

6. REFERENCES

[1] K. Min, Z. Zhang, J. Wright, and Y. Ma, "Decomposing background topics from keywords by principal component pursuit," in *CIKM*, 2010.

[2] S. Brutzer, B. Hoferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *CCVPR*, 2011, pp. 1937–1944.

[3] E.J. Candès, X. Li, and J. Ma, Y. andb Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, Article 11, 2011.

[4] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky, "Sparse and low-rank matrix decompositions," in *Sysid*, 2009.

[5] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan, "Multi-task low-rank affinity pursuit for image segmentation," in *ICCV*, 2011, pp. 2439–2446.

[6] Z. Zeng, T.H. Chan, K. Jia, and D. Xu, "Finding correspondence from multiple images via sparse and low-rank decomposition," in *ECCV*, 2012, pp. 325–339.

[7] F. Yang, H. Jiang, Z. Shen, W. Deng, and D.N. Metaxas, "Adaptive low rank and sparse decomposition of video using compressive sensing," *CoRR*, vol. abs/1302.1610, 2013.

[8] Y. Peng, A. Ganesh, J. Wright, and Y. Xu, W. andMa, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2233–2246, 2012.

[9] Z. Shi, J. Han, T. Zheng, and S. Deng, "Online learning for classification of low-rank representation features and its applications in audio segment classification," *CoRR*, vol. abs/1112.4243, 2011.

[10] Y.H. Yang, D. Bogdanov, P. Herrera, and M. Sordo, "Music retagging using label propagation and robust principal component analysis," in *WWW*, New York, NY, USA, 2012, pp. 869–876.

[11] W. Cai, Q. Li, and X. Guan, "Automatic singer identification based on auditory features," 2011.

[12] J. Salamon, E. Gómez, D.P.W. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications and challenges," *IEEE Signal Process. Mag.*, 2013.

[13] R.B. Dannenberg, W.P. Birmingham, B. Pardo, N. Hu, C. Meek, and G. Tzanetakis, "A comparative evaluation of search techniques for query-by-humming using the musart testbed," *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 5, pp. 687–701, 2007.

[14] B. Zhu, W. Li, R. Li, and X. Xue, "Multi-stage non-negative matrix factorization for monaural singing voice separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2096–2107, 2013.

[15] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," in *ICASSP*, 2011.

[16] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *ICASSP*, 2012.

[17] D. FitzGerald, "Vocal separation using nearest neighbours and median filtering," in *ISSC*, 2012.

[18] P.S. Huang, S.D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing voice separation from monaural recordings using robust principal component analysis," in *ICASSP*, 2012.

[19] C.L. Hsu and J.S.R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 310–319, 2010.

[20] Y.H. Yang, "On sparse and low-rank matrix decomposition for singing voice separation," in *MM*, 2012, pp. 757–760.

[21] M. Moussallam, G. Richard, and L. Daudet, "Audio source separation informed by redundancy with greedy multiscale decompositions," in *EUSIPCO*, 2012, pp. 2644–2648.

[22] S.G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Audio, Speech, Language Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[23] P. Sprechmann, A. Bronstein, and G. Sapiro, "Real-time online singing voice separation from monaural recordings using robust low rank modeling," in *ISMIR*, 2012.

[24] A. Lefévre, F. Glineur, and P.A. Absil, "A nuclear-norm based convex formulation for informed source separation," in *ESANN*, 2013.

[25] Y.H. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," in *ISMIR*, 2013.

[26] J. Salamon, *Melody Extraction from Polyphonic Music Signals*, Ph.D. thesis, Department of Information and Communication Technologies Universitat Pompeu Fabra, Barcelona, Spain, 2013.

[27] A.L. Berenzweig and D.P.W. Ellis, "Locating singing voice segments within music signals," in *WASPAA*, 2001, pp. 119–122.

[28] T.L. Nwe and Y. Wang, "Automatic detection of vocal segments in popular songs," in *Proc. ISMIR*, 2004, pp. 138–145.

[29] L. Feng, A.B. Nielsen, and L.K. Hansen, "Vocal segment classification in popular music.," in *ISMIR*, 2008, pp. 121–126.

[30] M. Fazel, *Matrix Rank Minimization with Applications*, Ph.D. thesis, Dept of Elec. Eng., Stanford Univ., 2002.

[31] B. Recht, M. Fazel, and P.A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[32] E.J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[33] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," Tech. Rep. UILU-ENG-09-2214, UIUC Tech. Rep., 2009.

[34] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," Tech. Rep. UILU-ENG-09-2215, UIUC, 2009.

[35] Xiaoming Yuan and Junfeng Yang, "Sparse and low-rank matrix decomposition via alternating direction methods," *Preprint*, pp. 1–11, 2009.

[36] J.F. Cai, E.J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[37] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[38] S. Chen, L. David, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1998.

[39] Z. Gao, L.F. Cheong, and M. Shan, *Block-Sparse RPCA for Consistent Foreground Detection*, vol. 7576 of *Lecture Notes in Computer Science*, pp. 690–703, Springer Berlin Heidelberg, 2012.

[40] Y. Grandvalet, "Least absolute shrinkage is equivalent to quadratic penalization," in *ICANN 98*, L. Niklasson, M. Boden, and T. Ziemke, Eds., Perspectives in Neural Computing, pp. 201–206. Springer London, 1998.

[41] H. Zou, "The adaptive lasso and its oracle properties," *J. Am. Statist. Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.

[42] D. Angelosante and G. Giannakis, "Rls-weighted lasso for adaptive estimation of sparse signals," in *ICASSP*, 2009, pp. 3245–3248.

[43] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, pp. 1759–1770, 2012.

[44] J.L. Durrieu, G. Richard, B. David, and C. Fevotte, ," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 564–575, March 2010.

[45] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[46] D. FitzGerald and M. Gainza, "Single channel vocal separation using median filtering and factorisation techniques," *ISAST Transactions on Electronic and Signal Processing*, vol. 4, no. 1, pp. 62–73, 2010.

[47] Z. Rafii, F. Germain, D.L. Sun, and G.J. Mysore, "Combining modeling of singing voice and background music for automatic separation of musical mixtures," in *ISMIR*, 2013.

[48] G. E. Poliner, D. P. W. Ellis, F. Ehmann, E. Gómez, S. Streich, and B. Ong, "Melody transcription from music audio: Approaches and evaluation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1247–1256, 2007.

[49] Z. Chen and D.P.W. Ellis, "Speech enhancement by sparse, low-rank and dictionary spectrogram decomposition," in *WASPAA*, 2013.