Combining Rhythm-Based and Pitch-Based Methods for Background and Melody Separation

Zafar Rafii, *Student Member, IEEE*, Zhiyao Duan, *Member, IEEE*, and Bryan Pardo, *Member, IEEE*

***Abstract*—Musical works are often composed of two characteristic components: the background (typically the musical accompaniment), which generally exhibits a strong rhythmic structure with distinctive repeating time elements, and the melody (typically the singing voice or a solo instrument), which generally exhibits a strong harmonic structure with a distinctive predominant pitch contour. Drawing from findings in cognitive psychology, we propose to investigate the simple combination of two dedicated approaches for separating those two components: a rhythm-based method that focuses on extracting the background via a rhythmic mask derived from identifying the repeating time elements in the mixture and a pitch-based method that focuses on extracting the melody via a harmonic mask derived from identifying the predominant pitch contour in the mixture. Evaluation on a data set of song clips showed that combining such two contrasting yet complementary methods can help to improve separation performance—from the point of view of both components—compared with using only one of those methods, and also compared with two other state-of-the-art approaches.**

***Index Terms*—Background, melody, pitch, rhythm, separation.**

I. Introduction

THE ability to separate a musical mixture into its background component (typically the musical accompaniment) and its melody component (typically the singing voice or a solo instrument) can be useful for many applications, e.g., karaoke gaming (need the background), query-by-humming (need the melody), or audio remixing (need both components). Existing methods for background and melody separation focus on modeling either the background (e.g., by learning a model from the non-vocal segments) or the melody (e.g., by identifying the predominant pitch contour), or both components concurrently (e.g., via joint or hybrid methods).

A. Melody-Focused Methods

Panning-based methods focus on modeling the melody by exploiting the inter-channel information in the mixture, assuming a two-channel mixture with a center-panned melody. Sofinanos *et al.* used a framework based on Independent Component Analysis (ICA) [1]. Kim *et al.* used a framework based on Gaussian Mixture Models (GMM) with inter-channel level differences and inter-channel phase differences [2].

Pitch-based methods focus on modeling the melody by identifying the predominant pitch contour in the mixture and inferring the harmonic structure of the melody. Meron *et al.* used prior pitch information to separate singing voice and piano accompaniment [3]. Zhang *et al.* used a framework based on a monophonic pitch detection algorithm [4]. Li *et al.* used a predominant pitch detection algorithm [5]. Hsu *et al.* used that same framework, additionally separating the unvoiced singing voice [6]. Hsu *et al.* then used a framework where singing pitch estimation and singing voice separation are performed jointly and iteratively [7]. Fujihara *et al.* also used a predominant pitch detection algorithm [8]. Cano *et al.* too [9], then additionally using prior information and additivity constraint [10]. Ryynänen *et al.* used a multi-pitch detection algorithm [11]. Lagrange *et al.* used a framework based on a graph partition problem [12].

Harmonic/percussive separation-based methods focus on modeling the melody by using a harmonic/percussive separation method on the mixture at different frequency resolutions, assuming the melody (typically the singing voice) as a harmonic component at low frequency resolution and a percussive component at high frequency resolution. FitzGerald *et al.* used a framework based on multiple median filters [13]. Tachibana *et al.* used a framework based on Maximum A Posteriori (MAP) estimation [14].

B. Background-Focused Methods

Adaptation-based methods focus on modeling the background by learning a model from the non-vocal segments in the mixture, which is then used to estimate the melody. Ozerov *et al.* used a framework based on GMM with Maximum Likelihood Estimation (MLE) [15] and MAP estimation [16]. Raj *et al.* used a framework based on Probabilistic Latent Component Analysis (PLCA) [17]. Han *et al.* also used PLCA [18].

Repetition or rhythm-based methods focus on modeling the background by identifying and extracting the repeating patterns in the mixture, assuming the background as a repeating component and the melody as a non-repeating component. Rafii *et al.* used a beat spectrum to first identify the periodically repeating patterns and a median filter to then extract the repeating background [19]. Liutkus *et al.* used a beat spectrogram to further identify the varying-periodically repeating patterns [20]. Rafii *et al.* then used a similarity matrix to also identify the non-periodically repeating patterns [21]. FitzGerald instead used a distance matrix [22].

Manuscript received January 22, 2014; revised May 26, 2014; accepted August 26, 2014. Date of publication September 04, 2014; date of current version September 16, 2014. This work was supported by the National Science Foundation (NSF) under Grant IIS-0812314. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. DeLiang Wang.

Z. Rafii and B. Pardo are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: zafarrafii@u.northwestern.edu; pardo@northwestern.edu).

Z. Duan is with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627-0126 USA (e-mail: zhiyao.duan@rochester.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2354242

C. Joint Methods

Non-negative Matrix Factorization (NMF)-based methods model both components concurrently by decomposing the mixture into non-negative elements and clustering them into background and melody. Vembu *et al.* used NMF (and also ICA) with trained classifiers and different features [23]. Chanrungutai *et al.* used NMF with rhythmic and continuous cues [24]. Zhu *et al.* used multiple NMFs at different frequency resolutions with spectral and temporal discontinuity cues [25]. Durrieu *et al.* used a framework based on GMM [26] and an Instantaneous Mixture Model (IMM) [27] with an unconstrained NMF model for the background and a source-filter model for the melody (typically the singing voice). Joder *et al.* used the same IMM framework, additionally exploiting an aligned musical score [28]. Marxer *et al.* used the same IMM framework, with a Tikhonov regularization instead of NMF [29]. Bosch *et al.* used that same framework, additionally exploiting a misaligned musical score [30]. Janer and Marxer used that same framework, additionally separating the unvoiced fricatives [31] and the voice breathiness [32].

Robust Principal Component Analysis (RPCA)-based methods model both components concurrently by decomposing the mixture into a low-rank component and a sparse component, assuming the background as low-rank and the melody as sparse. Huang *et al.* used a framework based on RPCA [33]. Sprechmann *et al.* also used RPCA, introducing a non-negative variant of RPCA and proposing two efficient feed-forward architectures [34]. Yang also used RPCA, including the incorporation of harmonicity priors and a back-end drum removal procedure [35]. Yang then used RPCA, computing the low-rank representations of both the background and the melody [36]. Papadopoulos *et al.* also used RPCA, incorporating music content information to guide the decomposition [37].

Very recently, Liutkus *et al.* used a framework based on local regression with proximity kernels, assuming that a component can be modeled through its regularities, e.g., periodicity for the background and smoothness for the melody [38].

D. Hybrid Methods

Hybrid methods model both components concurrently by combining different methods. Cobos *et al.* used a panning-based method and a pitch-based method [39]. Virtanen *et al.* used a pitch-based method to first identify the vocal segments of the melody and an adaptation-based method with NMF to then learn a model from the non-vocal segments for the background [40]. Wang *et al.* used a pitch-based method and an NMF-based method with a source-filter model [41]. FitzGerald used a repetition-based method to first estimate the background and a panning-based method to then refine background and melody [42]. Rafii *et al.* used an NMF-based method to first learn a model for the melody and a repetition-based method to then refine the background [43].

E. Motivating Psychological Research

Perceptual psychologists have been studying the ability of humans to attend to and process meaningful elements in the auditory scene for decades. In this literature, following the seminal work of Bregman [44], separation of the audio scene into meaningful elements is referred to as *streaming*. When humans focus attention on some part of the auditory scene they are performing streaming, as focus on one element necessarily requires parsing the scene into parts corresponding to that element and parts that do not correspond to it.

Studies have shown humans are able to easily focus on the background or the melody when listening to musical mixtures, by allocating their attention to either the rhythmic structure or the pitch structure [45], [46]. Recent work [47] in the Proceedings of the National Academy of Science has also documented human ability to isolate sounds based on regular repetition and treat these as unique perceptual units, and has even proposed that the human system could use a mechanism similar to that used in rhythm-based source separation methods.

Perceptual studies have shown that rhythm and melody are two essential dimensions in music processing, with the rhythmic dimension arising from temporal variations and repetitions and the melodic dimension arising from pitch variations [45], [48], [49]. Most studies have found that rhythm and melody are not treated jointly, but rather processed separately and then later integrated to produce a unified experience of the musical mixture [45], [46], [48]–[54]. In particular, some of those studies have suggested that rhythm and melody are processed by two separate subsystems and a simple additive model is sufficient to account for their independent contributions [46], [49]–[52]. These findings are supported by case studies of patients suffering from amusia, where some were found impaired in their processing of melody with preserved processing of rhythm (amelodia) [48], [50]–[52] and others were found impaired in their processing of rhythm with preserved processing of melody (arrhythmia) [50], [51], [53], [54].

F. Motivation and Rationale for our Approach

We take inspiration from the psychological literature (see Section I-E) to guide potential directions for our system development. We do not wish to perform cognitive modeling, where the goal is to exactly duplicate the mechanisms by which humans parse the auditory scene. Instead, we draw broad directions from this body of knowledge to guide our system design.

Since multiple studies indicate that humans use rhythm and pitch as independent elements that are then integrated to segment the audio scene into streams, we propose to use a simple combination of a rhythm-based and a pitch-based method to separate foreground from background. Since there is no broad agreement in the psychological literature about how rhythm and pitch based processing may be combined, we compare the two simplest approaches (serial and parallel combinations). While many other combinations are possible, exploring all possible combination methods would lengthen the work excessively and overwhelm the reader with experimental variations.

We are not performing cognitive modeling, therefore we favor the simplicity of using standard signal representations used in audio source separation (e.g., magnitude spectrograms), rather than a representation based on a faithful model of the ear [55] or auditory cortex [56].

This choice of a standard signal representation lets us use a standard approach to creating system output from both the rhythm and the pitch-based systems: time-frequency masking. Since both systems output time-frequency masks, this makes for a simple, modular approach to combining systems by combining masks. It also lets other researchers easily duplicate our combination work as it is simple to understand and replicate.

Our choices of systems for rhythm and pitch-based source separation approaches were pragmatic. We selected simple systems that have been published within the last few years, that showed good results in comparative studies, and to which we have access to the source code so we could ensure each system outputs a time-frequency mask in a compatible format. Since the focus of the study is to explore how a simple combination of simple rhythm and a pitch-based methods may affect source separation, we did not compare multiple pitch or repetition-based separation systems, although we are aware many excellent pitch-based and rhythm-based systems exist (see Sections Section I-A and Section I-B for an overview).

In testing our systems we focus on two questions. First: Is it better to combine rhythm and pitch-based methods for source separation in series or in parallel? How does the performance of a simple combination of rhythm and pitch separation compare to existing state-of-the-art systems that combine multiple approaches to source separation. Therefore, we separate our experimental into these two sections. Our choice of data sets and error measures were made to favor broadly-used data and error measures.

The rest of the article is organized as follows. In Section II, we describe the rhythm-based and the pitch-based method, and propose a parallel and a series combination of those two methods. In Section III, we analyze the parallel and the series combination on a data set of 1,000 song clips using different weighting strategies. In Section IV, we compare the rhythm-based and pitch-based methods, and the best of the parallel and series combinations with each other, and against two other state-of-the-art methods. In Section V, we conclude this article.

II. Methods

In this section, we describe the rhythm-based and the pitch-based method, and propose a parallel and a series combination of those two methods.

A. Rhythm-based Method

Studies in cognitive psychology (see Section I-E for the full overview) have shown that humans are able to focus on the background in musical mixtures by allocating their attention to the rhythmic structure that arises from the temporal variations [45], [46], [48], [49]. Drawing from these findings, we propose to extract the background by using a rhythm-based method that derives a rhythmic mask from identifying the repeating time elements in the mixture.

Assuming that the background is the predominant repeating component in the mixture, repetition-based methods typically first identify the repeating time elements by using a beat spectrum/spectrogram or a similarity/distance matrix, and then remove the non-repeating time elements by using a median filter at repetition rate [19]–[22] (see Section I-B).

In this work, we chose a repetition-based method that is referred to as REPET-SIM. REPET-SIM is a generalization of the REpeating Pattern Extraction Technique (REPET) [19] that uses a similarity matrix to identify the repeating elements of the background music [21].

The method can be summarized as follows. First, it identifies the repeating elements by computing a similarity matrix from the magnitude spectrogram of the mixture and locating the time frames that are the most similar to one another. Then, it derives a repeating model by median filtering the time frames of the magnitude spectrogram at their repetition rate. Finally, it extracts the repeating structure by refining the repeating model and deriving a rhythmic mask. For more details about the method, the reader is referred to [21].

B. Pitch-Based Method

Studies in cognitive psychology (see Section I-E for the full overview) have also shown that humans can focus on the melody in musical mixtures by attending to the pitch structure of the audio [45], [46], [48], [49]. Drawing from these findings, we chose to extract the melody by using a pitch-based method that derives a harmonic mask from identifying the predominant pitch contour in the mixture.

Assuming that the melody is the predominant harmonic component in the mixture, pitch-based methods typically first identify the predominant pitch contour by using a pitch detection algorithm, and then infer the corresponding harmonics by computing the integer multiples of the predominant pitch contour [3]–[12] (see Section I-A).

In this work, we chose a pitch-based method that will be referred to as *Pitch*. Pitch uses a multi-pitch estimation approach [57] to identify the pitch contour of the singing voice. Although originally proposed for multi-pitch estimation of general harmonic mixtures, the algorithm has been systematically evaluated for predominant pitch estimation and shown to work well compared with other melody extraction methods [18]. In this work, we modified the method in [57] to better suit it for melody extraction. While other excellent approaches to melody extraction exist (e.g., Hsu *et al.* [7]), the focus of this work is on combining a simple and clear pitch-based method with a simple and clear rhythm-based method, rather than a comparison of pitch-based methods for source separation. Therefore, we selected a known-good method for which we have a deep understanding of the inner workings and access to the source code.

The method can be summarized as follows. First, it identifies peaks in every spectrum of the magnitude spectrogram of the mixture using the method in [58], also defining non-peak regions, and estimates the predominant pitch using the method in [57], from the peaks and non-peak regions. Then, it forms pitch contours by connecting pitches that are close in time (in adjacent frames) and frequency (difference less than 0.3 semitone). Small time gaps (less than 100 milliseconds) between two successive pitch contours are filled with their average pitch value so that the two contours are merged into a longer one, if their pitch difference is small (less than 0.3 semitone). Shorter pitch contours (less than 100 milliseconds) are removed. This is to remove some musical noise caused by pitch detection errors in individual frames [59].

Since some estimated pitches may actually correspond to the accompaniment instead of the melody, we used a simple method to discriminate pitch contours of melody and accompaniment, assuming that melody pitches vary more (due to vibratos) than accompaniment pitches [60]. More specifically, we calculated the pitch variance for each pitch contour, and removed the ones whose variance is less than 0.05 square semitones. The remaining pitch contours are supposed to be

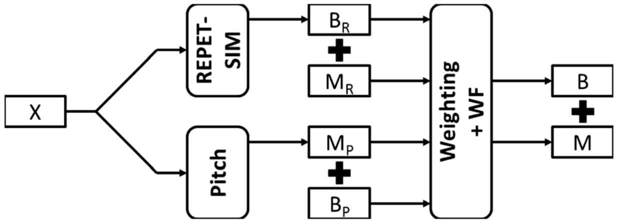


Fig. 1. Diagram of the parallel combination (see Section II-C).

those of the melody. Finally, we computed a harmonic mask to extract the melody. All the thresholds in this algorithm are set through observation of several songs. No optimization was performed to tune them.

C. Parallel Combination

Studies in cognitive psychology have further shown that humans process rhythm and melody separately to then later integrate them in order to produce a unified experience of the musical mixture [45], [46], [48]–[54]. Drawing from these findings, we propose to separate the background and the melody by using a parallel combination of the rhythm-based method and the pitch-based method.

The method can be summarized as follows. Given a mixture spectrogram X, REPET-SIM derives a background mask $B_R(\in [0,1])$—and the complementary melody mask $M_R(=\mathbb{1}-B_R)$, and Pitch derives a melody mask $M_P(\in [0,1]))$—and the complementary background mask $B_P(=\mathbb{1}-M_P)$, concurrently. The final background mask $B(\in [0,1])$ and the final melody mask $M(\in [0,1])$ are then derived by weighting and Wiener filtering (WF) the masks B_R, M_R, M_P, and B_P, appropriately so that $B+M=1$ (see Fig. 1). Here, 1 represents a matrix of all ones.

We use two weight parameters, w_B and $w_M(\in [0,1])$, when combining the background masks, B_R and B_P, and the melody masks, M_R and M_P, obtained from REPET-SIM and Pitch, respectively (see Equation (1)). We will analyze the separation performance using different values of w_B and w_M for deriving the final background mask B and the final melody mask M (see Section III-D). Here, $A\circ B$ and $\cdot\frac{A}{B}$ represent the element-wise multiplication and the element-wise division, respectively, between the matrices A and B.

$$
\begin{aligned}
\textbf{REPET-SIM}&: X=(B_R+M_R)\circ X, \quad \{B_R,M_R\}\in[0,1]\\
\textbf{Pitch}&: X=(B_P+M_P)\circ X, \quad \{B_P,M_P\}\in[0,1]\\
\textbf{Weighting}&: \begin{cases} B\leftarrow w_B B_R+(1-w_B)B_P, & w_B\in[0,1]\\ M\leftarrow w_M M_R+(1-w_M)M_P, & w_M\in[0,1]\end{cases}\\
\textbf{WF}&: B\leftarrow\cdot\frac{B}{B+M} \quad\text{and}\quad M\leftarrow\cdot\frac{M}{B+M}\\
\textbf{Parallel}&: X=(B+M)\circ X, \quad \{B,M\}\in[0,1] \qquad (1)
\end{aligned}
$$

Since REPET-SIM focuses on extracting the background and Pitch focuses on extracting the melody, we hypothesize that the best separation performance will be obtained when the final background mask is derived by mostly using the background mask from REPET-SIM (i.e., $w_B\approx 1$) and the final melody mask is derived by mostly using the melody mask from Pitch (i.e., $w_M\approx 0$) (see Section III-D).

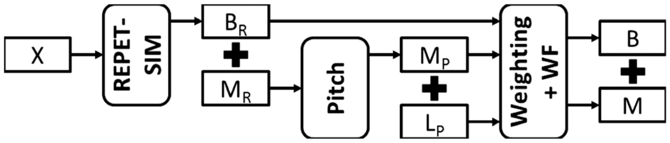


Fig. 2. Diagram of the series combination (see Section II-D).

D. Series Combination

Additionally, a musical mixture can be understood as the sum of a pitched melody, a repeating background, and an extra component comprising the non-repeating pitched elements of the background. On this basis, we also propose to separate the background and the melody by using a series combination of the rhythm-based method and the pitch-based method. Since REPET-SIM is more robust than Pitch when directly applied on a mixture, we chose to first use REPET-SIM to separate the components, and then Pitch to refine the estimates.

The method can be summarized as follows. Given a mixture spectrogram X, REPET-SIM first derives a background mask $B_R(\in [0,1])$—and the complementary melody mask $M_R(=\mathbb{1}-B_R)$. Given the melody mask M_R, Pitch then derives a refined melody mask M_P $(\leq M_R)$—and a complementary "leftover" mask $L_P(=M_R-M_P)$. The final background mask $B(\in [0,1])$ and the final melody mask $M(\in [0,1])$ are then derived by weighting and Wiener filtering (WF) the masks B_R, M_P, and L_P, appropriately so that $B+M=1$ (see Fig. 2). Here, $\mathbb{1}$ represents a matrix of all ones.

We use a weight parameter, $w(\in [0,1])$, when refining the background mask, B_R, and the melody mask, M_P, obtained from REPET-SIM and Pitch, respectively (see Equation (2)). We will analyze the separation performance using different values of w for deriving the final background mask B and the final melody mask M (see Section III-E). Here, $A\circ B$ and $\cdot\frac{A}{B}$ represent the element-wise multiplication and the element-wise division, respectively, between the matrices A and B.

$$
\begin{aligned}
\textbf{REPET-SIM}&: X=(B_R+M_R)\circ X, \quad \{B_R,M_R\}\in[0,1]\\
\textbf{Pitch}&: M_R=M_P+L_P, \quad \{M_P,L_P\}\leq M_R\\
\textbf{Weighting}&: \begin{cases} B\leftarrow B_R+wL_P,\\ M\leftarrow M_P+(1-w)L_P,\end{cases} \quad w\in[0,1]\\
\textbf{WF}&: B\leftarrow\cdot\frac{B}{B+M} \quad\text{and}\quad M\leftarrow\cdot\frac{M}{B+M}\\
\textbf{Series}&: X=(B+M)\circ X, \quad \{B,M\}\in[0,1] \qquad (2)
\end{aligned}
$$

Since REPET-SIM focuses on extracting the repeating background and Pitch focuses on extracting the pitched melody, the extra leftover is most likely to comprise the non-repeating pitched elements of the background, so we hypothesize that the best separation performance will be obtained when the final background mask and the final melody mask are derived by mostly adding the leftover mask from Pitch to the background mask from REPET-SIM (i.e., $w\approx 1$) (see Section III-E).

III. Evaluation 1

In this section, we analyze the parallel and the series combination on a data set of 1,000 song clips using different weighting strategies.

A. Data Set

The MIR-1K[1] dataset consists of 1,000 song clips in the form of split stereo WAVE files sampled at 16 kHz, with the background and melody components recorded on the left and right channels, respectively. The song clips were extracted from 110 karaoke Chinese pop songs performed by amateur singers consisting of 8 females and 11 males. The duration of the clips ranges from 4 to 13 seconds [6].

We then derived a set of 1,000 mixtures by summing, for each song clip, the left channel (i.e., the background) and the right channel (i.e., the melody) into a monaural mixture.

B. Performance Measures

The BSS Eval[2] toolbox consists of a set of measures that intend to quantify the quality of the separation between a source and its estimate. The principle is to decompose an estimate into contributions corresponding to the target source, the interference from unwanted sources, and the artifacts such as “musical noise.” Based on this principle, the following measures were then defined (in dB): Source to Interference Ratio (SIR), Sources to Artifacts Ratio (SAR), and Signal to Distortion Ratio (SDR) which measures the overall error [61]. We chose 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 [62]. These measures are broadly used in the source separation community.

We then derived three measures, that will be referred to as Δ SIR, Δ SAR, and Δ SDR, by taking the difference between the SIR, SAR, and SDR computed using the estimated masks from a given method, and the SIR, SAR, and SDR computed using the ideal masks from the original sources, respectively. Δ SIR, Δ SAR, and Δ SDR basically measure how close the separation performance can get to the maximal separation performance given a masking approach. Values are logically negative (i.e., ≤ 0), with higher values (i.e., closer to 0) meaning better separation performance.

C. Algorithm Parameters

Given the REPET-SIM algorithm[3], we used Hamming windows of 1024 samples, corresponding to 64 milliseconds at a sampling frequency of 16 kHz, with an overlap of 50%. The minimal threshold between similar frames was set to 0, the minimal distance between consecutive frames to 0.1 seconds, and the maximal number of repeating frames to 50 [21].

Given the Pitch algorithm[4], we used Hamming windows of 512 samples, corresponding to 32 milliseconds at a sampling frequency of 16 kHz, with an overlap of 75%. The predominant pitch was estimated between 80 and 600 Hz, and the minimal time and pitch differences for merging successive pitches were set to 100 milliseconds and 0.3 semitones, respectively [57], [58].

The masks for REPET-SIM and Pitch were then derived from their corresponding estimates, by using the same parameters that we used for REPET-SIM, i.e., Hamming windows of 1024 samples, corresponding to 64 milliseconds at a sampling frequency of 16 kHz, with an overlap of 50%.

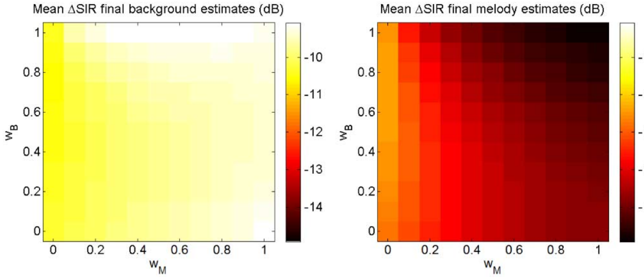


Fig. 3. Mean Δ SIR for the final background estimates (left plot) and the final melody estimates (right plot), for the parallel combination for different weights w_B and w_M. Lighter values are better (see Section III-D).

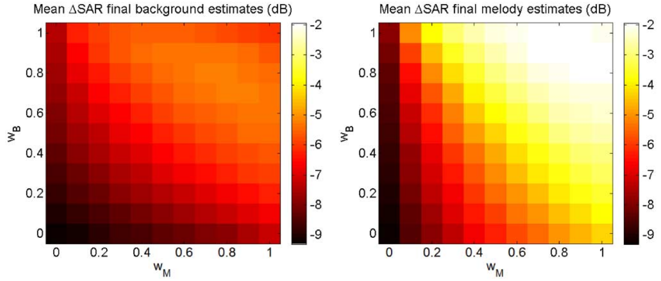


Fig. 4. Mean Δ SAR for the final background estimates (left plot) and the final melody estimates (right plot), for the parallel combination for different weights w_B and w_M. Lighter values are better (see Section III-D).

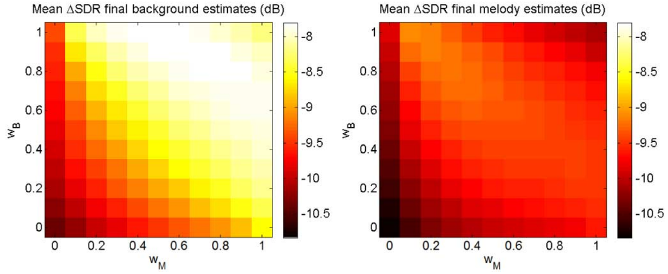


Fig. 5. Mean Δ SDR for the final background estimates (left plot) and the final melody estimates (right plot), for the parallel combination for different weights w_B and w_M. Lighter values are better (see Section III-D).

D. Parallel Combination

Fig. 3, Fig. 4 and Fig. 5 show the mean Δ SIR, mean Δ SAR, and mean Δ SDR, respectively, for the final background estimates (left plot) and the final melody estimates (right plot), for the parallel combination for different weights w_B and w_M (from 0 to 1 in steps of 0.1). Lighter values are better.

Fig. 3 suggests that, for less interference in the final background estimates, the background mask from REPET-SIM, B_R, should be weighted more than the background mask from Pitch, B_P, and the melody mask from REPET-SIM, M_R, and the melody mask from Pitch, M_P, should be weighted equally, when deriving the final background mask, B; for less interference in the final melody estimates, B_R and B_P should be weighted equally, and M_R should be weighted less than M_P, when deriving the final melody mask, M.

Fig. 4 suggests that, for less artifacts in the final background estimates and the final melody estimates, B_R and M_R should

[1] http://sites.google.com/site/unvoicedsoundseparation/mir-1k

[2] http://bass-db.gforge.inria.fr/bss_eval/

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

[4] http://music.eecs.northwestern.edu/research.php?project=mpitch

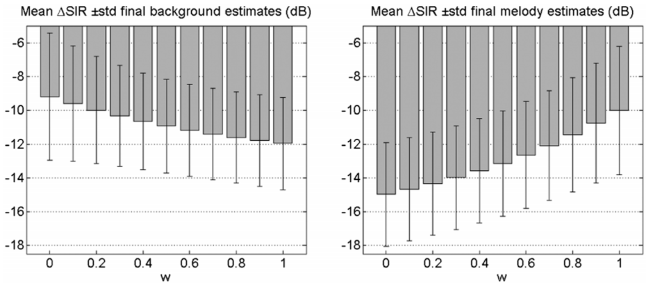


Fig. 6. Mean Δ SIR ± standard deviation for the final background estimates (left plot) and the final melody estimates (right plot), for the series combination for different weights w. Higher values are better (see Section III-E).

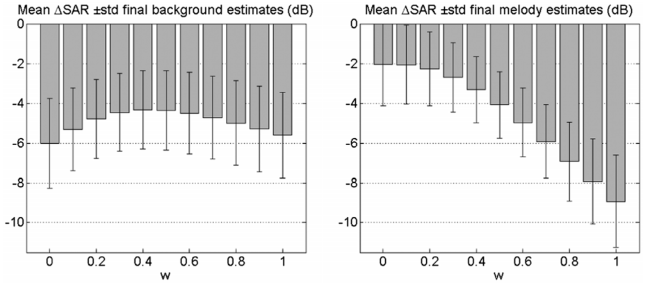


Fig. 7. Mean Δ SAR ± standard deviation for the final background estimates (left plot) and the final melody estimates (right plot), for the series combination for different weights w. Higher values are better (see Section III-E).

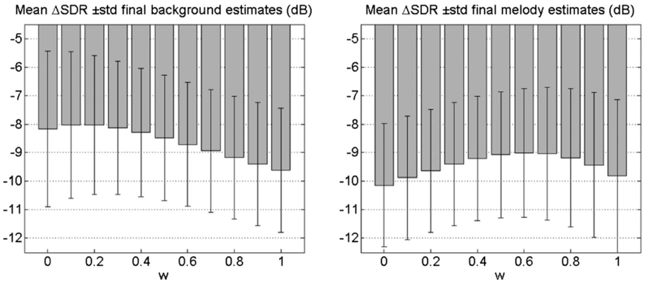


Fig. 8. Mean Δ SDR ± standard deviation for the final background estimates (left plot) and the final melody estimates (right plot), for the series combination for different weights w. Higher values are better (see Section III-E).

be weighted more than B_P and M_P, when deriving B and M, respectively.

Fig. 5 suggests that, for less overall error in the final background estimates, B_R should be weighted more than B_P, and M_R and M_P should be weighted equally when deriving B; for less overall error in the final melody estimates, B_R should be weighted more than B_P, and M_R and M_P should be weighted equally, when deriving M.

The results for the parallel combination show that the best separation performance is obtained when the final background mask is derived by using mostly the background mask from REPET-SIM, and the final melody mask is derived by mixing part of the melody mask from REPET-SIM with the melody mask from Pitch. While the results for the Δ SIR support our hypothesis (see Section II-C), the results for the Δ SAR do not, probably because Pitch tends to introduce musical noise in its estimates; this can be reduced by compensating with the estimates of REPET-SIM, hence the results for the Δ SDR.

The best parallel combination given the highest mean Δ SDR averaged over the final background estimates and the final melody estimates is obtained for w_B of 1 and w_M of 0.3.

E. Series Combination

Figs 6, Fig 7 and Fig 8show the mean Δ SIR ± standard deviation, mean Δ SAR ± standard deviation, and mean Δ SDR ± standard deviation, respectively, for the final background estimates (left plot) and the final melody estimates (right plot), for the series combination for different weights w (from 0 to 1 in steps of 0.1). Higher values are better.

Fig. 6 suggests that, for less interference in the final background estimates, the leftover mask, L_P, should be weighted less with the background mask from REPET-SIM, B_R, and more with the melody mask from Pitch, M_P, when deriving the final background mask, B; for less interference in the final melody estimates, L_P should be weighted more with B_R and less with M_P, when deriving the final melody mask, M.

Fig. 7 suggests that, for less artifacts in the final background estimates, L_P should be weighted equally with B_R and M_P, when deriving B; for less artifacts in the final melody estimates, L_P should be weighted less with B_R and more with M_P, when deriving M.

Fig. 8 suggests that, for less overall error in the final background estimates, L_P should be weighted less with B_R and more with M_P, when deriving B; for less overall error in the final melody estimates, L_P should be weighted equally with B_R and M_P, when deriving M.

The results for the series combination show that the best separation performance is obtained when the final background mask and the final melody mask are derived by dividing the leftover mask equally between the background mask from REPET-SIM and the melody mask from Pitch. Rather than supporting our hypothesis (see Section II-D), the results for the Δ SIR show that the leftover seems to represent an extra component that would hurt both the final background estimates if added to the background estimates from REPET-SIM, and the final melody estimates if added to the melody estimates from Pitch, hence the results for the Δ SDR.

The best series combination given the highest mean Δ SDR averaged over the final background estimates and the final melody estimates is obtained for w of 0.4.

IV. Evaluation 2

In this section, we compare the rhythm-based and pitch-based methods, and the best of the parallel and series combinations with each other, and against two other state-of-the-art methods.

A. Competitive Methods

Durrieu *et al.* proposed a joint method for background and melody separation based on an NMF framework (see Section I-C). They used an unconstrained NMF model for the background and a source-filter model for the melody, and derived the estimates jointly in a formalism similar to the NMF algorithm. They also added a white noise spectrum to the melody model to better capture the unvoiced components [27]. Given the algorithm[5], we used an analysis window of 64

[5] http://www.durrieu.ch/research/jstsp2010.html

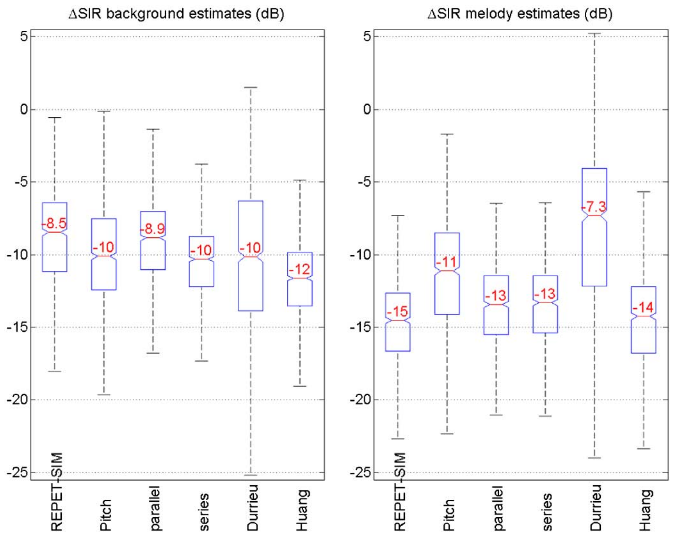


Fig. 9. Distribution of the Δ SIR for the background estimates (left plot) and the melody estimates (right plot), for REPET-SIM, Pitch, the best parallel combination, the best series combination, the method of Durrieu *et al.*, and the method of Huang *et al.* High values are better (see Section IV-B).

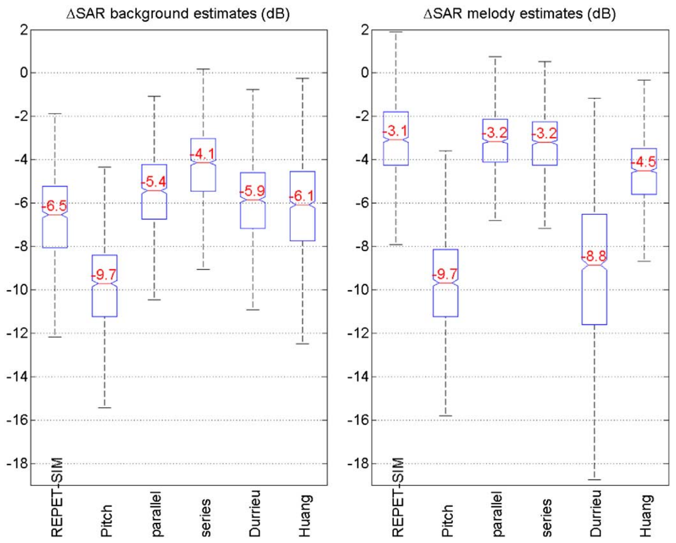


Fig. 10. Distribution of the Δ SAR for the background estimates (left plot) and the melody estimates (right plot), for REPET-SIM, Pitch, the best parallel combination, the best series combination, the method of Durrieu *et al.*, and the method of Huang *et al.*, High values are better (see Section IV-B).

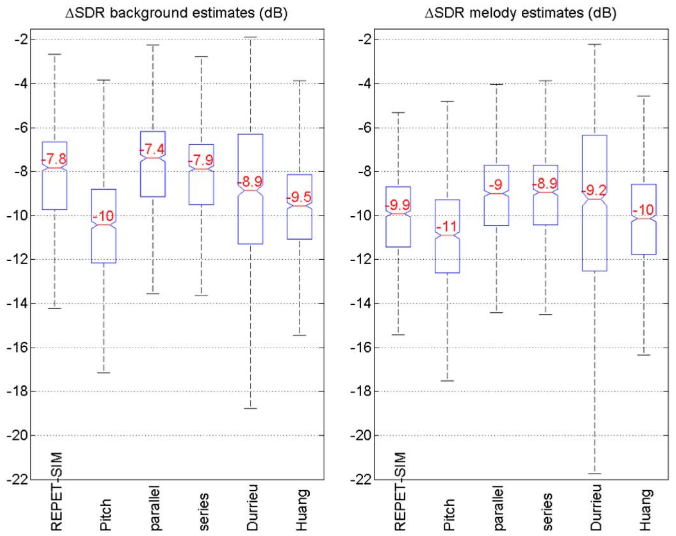


Fig. 11. Distribution of the Δ SDR for the background estimates (left plot) and the melody estimates (right plot), for REPET-SIM, Pitch, the best parallel combination, the best series combination, the method of Durrieu *et al.*, and the method of Huang *et al.*, High values are better (see Section IV-B).

milliseconds, an analysis Fourier size of 1024 samples, a step size of 32 milliseconds, and a number of 30 iterations.

Huang *et al.* proposed a joint method for background and melody separation based on an RPCA framework (see Section I-C). They used a low-rank model for the background and a sparse model for the melody, and derived the estimates jointly by minimizing a weighted combination of the nuclear norm and the L_1 norm. They assumed that, in musical mixtures, the background can be regarded as a low-rank component and the melody as a sparse component [33]. Given the algorithm[6], we used the default parameters.

B. Comparative Analysis

Fig. 9, Fig. 10 and Fig. 11 show the distribution of the Δ SIR, Δ SAR, and Δ SDR, respectively.

Recall that Δ SDR is an overall performance measure that combines degree of source separation (Δ SIR) with quality of the resulting signals (Δ SAR). Therefore, readers interested in a synopsis of overall separation performance should focus on the Δ SDR plot in Fig. 11. Readers interested specifically in how completely the background and foreground were separated should focus on the Δ SIR plot in Fig. 9. Readers interested specifically in how many artifacts were introduced into the separated signals by the source separation algorithm should focus on the Δ SAR plot in Fig. 10.

Each figure shows the background estimates (left plot) and the melody estimates (right plot), for REPET-SIM, Pitch, the best parallel combination of REPET-SIM and Pitch, i.e., for w_B of 1 and w_M of 0.3 (see Section III-C), the best series combination of REPET-SIM and Pitch, i.e., for w of 0.4 (see Section III-D), the method of Durrieu *et al.*, and the method of Huang *et al.* On each box, the central mark is the median (whose value is displayed in the box), the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data points not considered outliers (which are not shown here). Higher values are better.

Fig. 9 suggests that, for reducing the interference in the background estimates, the parallel combination and the series combination, when properly weighted, can perform as well or better than REPET-SIM and Pitch alone, and the competitive methods, although REPET-SIM seems still better than the series combination; for reducing the interference in the melody estimates, the method of Durrieu *et al.* still performs better than the other methods, although it shows a very large statistical dispersion, which means that, while it can do much better in some cases, it also does much worse in other cases.

[6] https://sites.google.com/site/singingvoiceseparationrpca/

Fig. 10 suggests that, for reducing the artifacts in the background estimates and the melody estimates, the parallel combination and the series combination, when properly weighted, can perform as well or better than REPET-SIM and Pitch alone, and the competitive methods, with the series combination performing better than the parallel combination for the background estimates.

Fig. 11 suggests that, for reducing the overall error in the background estimates and the melody estimates, the parallel combination and the series combination, when properly weighted, can overall perform better than REPET-SIM or Pitch alone, and the competitive methods, with the parallel combination performing slightly better than the series combination.

The results of the comparative analysis show that, when properly weighted, the parallel and the series combinations of a rhythm-based and a pitch-based method can, as expected, perform better than the rhythm-based or the pitch-based method alone, for background and melody separation. Furthermore, a combination of simple approaches can also perform better than (or at least as well as) state-of-the-art methods based on sophisticated approaches that jointly model the background and the melody.

C. Statistical Analysis

Since $\Delta\mathrm{SDR}$ is an overall measure of system performance that combines $\Delta\mathrm{SIR}$ and $\Delta\mathrm{SAR}$, we focus our statistical analysis on $\Delta\mathrm{SDR}$. We used a (parametric) analysis of variance (ANOVA) when the distributions were all normal, and a (nonparametric) Kruskal-Wallis test when one of the distributions was not normal. We used a Jarque-Bera test to determine if a distribution was normal or not.

For the $\Delta\mathrm{SIR}$ for the background estimates, the statistical analysis showed that REPET-SIM $=$ parallel $>$ Pitch $=$ Durrieu $>$ series $>$ Huang, where "$a = b$" means that a and b are not significantly different, and "$a > b$" means that a is significantly higher than b for the melody estimates, Durrieu $>$ REPET-SIM $>$ parallel $=$ series $>$ Pitch $=$ Huang.

For the $\Delta\mathrm{SAR}$ for the background estimates, the statistical analysis showed that series $>$ parallel $=$ Durrieu and Durrieu $=$ Huang, but parallel $>$ Huang, Huang $>$ REPET-SIM $>$ Pitch for the melody estimates, REPET-SIM $=$ parallel $=$ series $>$ Huang $>$ Durrieu $>$ Pitch.

For the $\Delta\mathrm{SDR}$ for the background estimates, the statistical analysis showed that parallel $>$ series $=$ REPET-SIM $>$ Durrieu $>$ Huang $>$ Pitch for the melody estimates, series $=$ parallel, and parallel $=$ Durrieu, but series $>$ Durrieu, Durrieu $>$ REPET-SIM $=$ Huang $>$ Pitch.

V. Conclusion

Inspired by findings in cognitive psychology, we investigated the simple combination of two dedicated approaches for separating background and melody in musical mixtures: a rhythm-based method that focuses on extracting the background by identifying the repeating time elements and a pitch-based method that focuses on extracting the melody by identifying the predominant pitch contour. Evaluation on a data set of song clips showed that a simple parallel and series combination, when properly weighted, can perform better than the rhythm-based or the pitch-based method alone, but also two other state-of-the-art methods based on more sophisticated approaches.

The separation performance of such combinations of course depends on how the rhythm-based method and the pitch-based method are combined, and on their individual separation performance regarding both the background component and the melody component. Given the findings in cognitive psychology and the results obtained here, we believe that further advancement in separating background and melody potentially lies in independently improving the analysis of the rhythm structure and the pitch structure in musical mixtures.

More information, including source codes and audio examples, can be found online.

Acknowledgment

The authors would like to thank Richard Ashley for his expertise in music cognition, Antoine Liutkus for his suggestion in using delta metrics, and the anonymous reviewers for their helpful comments on the article.

References

[1] S. Sofianos, A. Ariyaeeinia, and R. Polfreman, "Singing voice separation based on non-vocal independent component subtraction," in *Proc. 13th Int. Conf. Digital Audio Effects*, Graz, Austria, Sep. 6–10, 2010.

[2] M. Kim, S. Beack, K. Choi, and K. Kang, "Gaussian mixture model for singing voice separation from stereophonic music," in *Proc. AES 43rd Int. Conf.: Audio for Wirelessly Netw. Personal Devices*, Pohang, Korea, Sep.–Oct. 1–29, 2011, pp. 6–2.

[3] Y. Meron and K. Hirose, "Separation of singing and piano sounds," in *Proc. 5th Int. Conf. Spoken Lang. Process.*, Sydney, Australia, Nov. –Dec. 4–30, 1998.

[4] Y.-G. Zhang and C.-S. Zhang, "Separation of voice and music by harmonic structure stability analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, Amsterdam, Netherlands, Jul. 6–8, 2005, pp. 562–565.

[5] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1475–1487, May 2007.

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

[7] C.-L. Hsu, D. Wang, J.-S. R. Jang, and K. Hu, "A tandem algorithm for singing pitch extraction and voice separation from music accompaniment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1482–1491, Jul. 2012.

[8] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 638–648, Mar. 2010.

[9] E. Cano, C. Dittmar, and G. Schuller, "Efficient implementation of a system for solo accompaniment separation in polyphonic music," in *Proc. 20th Eur. Signal Process. Conf.*, Bucharest, Romania, Aug. 27–31, 2012, pp. 285–289.

[10] E. Cano, C. Dittmar, and G. Schuller, "Re-thinking sound separation: Prior information and additivity constraints in separation algorithms," in *Proc. 16th Int. Conf. Digital Audio Effects*, Maynooth, Ireland, Sep. 2–4, 2013.

[11] M. Ryynänen, T. Virtanen, J. Paulus, and A. Klapuri, "Accompaniment separation and karaoke application based on automatic melody transcription," in *Proc. IEEE Int. Conf. Multimedia Expo*, Hannover, Germany, Jun. 23–26, 2008, pp. 1417–1420.

[12] M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis, "Normalized cuts for predominant melodic source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 278–290, Feb. 2008.

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

[13] D. FitzGerald and M. Gainza, "Single channel vocal separation using median filtering and factorisation techniques," *ISAST Trans. Electron. Signal Process.*, vol. 4, no. 1, pp. 62–73, 2010.

[14] H. Tachibana, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 228–237, Jan. 2014.

[15] A. Ozerov, P. Philippe, and F. B. Rémi Gribonval, "One microphone singing voice separation using source-adapted models," in *IEEE Workshop Applicat. Signal Process. Audio Acoust.*. New Paltz, NY, USA: , Oct. 16–19, 2005, pp. 90–93.

[16] 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 Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1564–1578, Jul. 2007.

[17] B. Raj, P. Smaragdis, M. Shashanka, and R. Singh, "Separating a foreground singer from background music," in *Proc. Int. Symp. Frontiers Res. Speech Music*, Mysore, India, May 8–9, 2007.

[18] J. Han and C.-W. Chen, "Improving melody extraction using probabilistic latent component analysis," in *Proc. 36th Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, May 22–27, 2011, pp. 33–36.

[19] Z. Rafii and B. Pardo, "Repeating Pattern Extraction Technique (REPET): A simple method for music/voice separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 71–82, Jan. 2013.

[20] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *Proc. 37th Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 25–30, 2012, pp. 53–56.

[21] Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix," in *Proc. 13th Int. Soc. Music Inf. Retrieval*, Porto, Portugal, Oct. 8–12, 2012.

[22] D. FitzGerald, "Vocal separation using nearest neighbours and median filtering," in *Proc. 23nd IET Irish Signals Syst. Conf.*, Maynooth, Ireland, Jun. 28–29, 2012.

[23] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *Proc. 6th Int. Conf. Music Inf. Retrieval*, London, U.K., Sep. 11–15, 2005, pp. 337–344.

[24] A. Chanrungutai and C. A. Ratanamahatana, "Singing voice separation for mono-channel music using non-negative matrix factorization," in *Proc. Int. Conf. Adv. Technol. Commun.*, Hanoi, Vietnam, Oct. 6–9, 2008, pp. 243–246.

[25] B. Zhu, W. Li, R. Li, and X. Xue, "Multi-stage non-negative matrix factorization for monaural singing voice separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2093–2107, Oct. 2013.

[26] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 564–575, Mar. 2010.

[27] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1180–1191, Oct. 2011.

[28] C. Joder and B. Schuller, "Score-informed leading voice separation from monaural audio," in *Proc. 13th Int. Soc. Music Inf. Retrieval*, Porto, Portugal, Oct. 8–12, 2012.

[29] R. Marxer and J. Janer, "A Tikhonov regularization method for spectrum decomposition in low latency audio source separation," in *Proc. 37th Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 25–30, 2012, pp. 277–280.

[30] J. J. Bosch, K. Kondo, R. Marxer, and J. Janer, "Score-informed and timbre independent lead instrument separation in real-world scenarios," in *Proc. 20th Eur. Signal Process. Conf.*, Bucharest, Romania, Aug. 27–31, 2012, pp. 2417–2421.

[31] J. Janer and R. Marxer, "Separation of unvoiced fricatives in singing voice mixtures with semi-supervised NMF," in *Proc. 16th Int. Conf. Digital Audio Effects*, Maynooth, Ireland, Sep. 2–5, 2013.

[32] R. Marxer and J. Janer, "Modelling and separation of singing voice breathiness in polyphonic mixtures," in *Proc. 16th Int. Conf. Digital Audio Effects*, Maynooth, Ireland, Sep. 2–5, 2013.

[33] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa–Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. 37th Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 25–30, 2012, pp. 57–60.

[34] P. Sprechmann, A. Bronstein, and G. Sapiro, "Monaural recordings using robust low-rank modeling," in *Proc. 13th Int. Soc. Music Inf. Retrieval*, Porto, Portugal, Oct. 8–12, 2012.

[35] Y.-H. Yang, "On sparse and low-rank matrix decomposition for singing voice separation," in *Proc. 20th ACM Int. Conf. Multimedia*, Nara, Japan, Oct.–Nov. 2–29, 2012, pp. 757–760.

[36] Y.-H. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," in *Proc. 14th Int. Soc. Music Inf. Retrieval*, Curitiba, Brazil, Nov. 4–8, 2013.

[37] H. Papadopoulos and D. P. Ellis, "Music-content-adaptive robust principal component analysis for a semantically consistent separation of foreground and background in music audio signals," in *Proc. 17th Int. Conf. Digital Audio Effects*, Erlangen, Germany, Sep. 1–5, 2014.

[38] A. Liutkus, Z. Rafii, B. Pardo, D. FitzGerald, and L. Daudet, "Kernel spectrogram models for source separation," in *Proc. 4th Joint Workshop Hands-Free Speech Commun. Microphone Arrays*, Nancy, France, May 12–14, 2014.

[39] M. Cobos and J. J. López, "Singing voice separation combining panning information and pitch tracking," in *Proc. 124th Audio Eng. Soc. Conv.*, Amsterdam, The Netherlands, May 17–20, 2008, p. 7397.

[40] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music," in *Proc. ISCA Tutorial and Res. Workshop Statist. Percept. Audition*, Brisbane, Australia, Sep. 21, 2008, pp. 17–20.

[41] Y. Wang and Z. Ou, "Combining HMM-based melody extraction and NMF-based soft masking for separating voice and accompaniment from monaural audio," in *Proc. 36th Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, May 22–27, 2011, pp. 1–4.

[42] D. FitzGerald, "Stereo vocal extraction using adress and nearest neighbours median filtering," in *Proc. 16th Int. Conf. Digital Audio Effects*, Maynooth, Ireland, Sep. 2–4, 2013.

[43] Z. Rafii, D. L. Sun, F. G. Germain, and G. J. Mysore, "Combining modeling of singing voice and background music for automatic separation of musical mixtures," in *Proc. 14th Int. Soc. Music Inf. Retrieval*, Curitiba, PR, Czech Republic, Nov. 4–8, 2013.

[44] A. S. Bregman, *Auditory Scene Analysis*. Cambridge MA, USA: MIT Press, 1990.

[45] C. B. Monahan and E. C. Carterette, "Pitch and duration as determinants of musical space," *Music Percept.*, vol. 3, pp. 1–32, 1985, Fall.

[46] C. Palmer and C. L. Krumhansl, "Independent temporal and pitch structures in determination of musical phrases," *J. Experiment. Psychol.: Human Percept. Perform.*, vol. 13, no. 1, pp. 116–126, Feb. 1987.

[47] J. H. McDermott, D. Wrobleski, and A. J. Oxenham, "Recovering sound sources from embedded repetition," in *Proc. Natural Acad. Sci. United States of Amer.*, Jan. 18, 2011, vol. 108, no. 3, pp. 1188–1193.

[48] I. Peretz and R. Kolinsky, "Boundaries of separability between melody and rhythm in music discrimination: A neuropsychological perspective," *Quaterly J. Experiment. Psychol.*, vol. 46, no. 2, pp. 301–325, May 1993.

[49] C. L. Krumhansl, "Rhythm and pitch in music cognition," *Psychol. Bull.*, vol. 126, no. 1, pp. 159–179, Jan. 2000.

[50] I. Peretz, "Processing of local and global musical information by unilateral brain-damaged patients," *Brain*, vol. 113, no. 4, pp. 1185–1205, Aug. 1990.

[51] M. Schuppert, T. F. Münte, B. M. Wieringa, and E. Altenmüller, "Receptive amusia: Evidence for cross-hemispheric neural networks underlying music processing strategies," *Brain*, vol. 153, no. 3, pp. 546–559, Mar. 2000.

[52] M. Piccirilli, T. Sciarma, and S. Luzzi, "Modularity of music evidence from a case of pure amusia," *J. Neurol., Neurosurgery Psychiatry*, vol. 69, no. 4, pp. 541–545, Oct. 2000.

[53] M. D. Pietro, M. Laganaro, B. Leemann, and A. Schnider, "Receptive amusia: Temporal auditory processing deficit in a professional musician following a left temporo-parietal lesion," *Neuropsychologia*, vol. 42, no. 7, pp. 868–877, 2004.

[54] J. Phillips-Silver, P. Toiviainen, N. Gosselin, O. Piché, S. Nozaradan, C. Palmer, and I. Peretz, "Born to dance but beat deaf: A new form of congenital amusia," *Neuropsychologia*, vol. 49, no. 5, pp. 961–969, Apr. 2011.

[55] R. D. Patterson, "Auditory images. How complex sounds are represented in the auditory system," *J. Acoust. Soc. Jpn. (E)*, vol. 21, no. 4, pp. 183–190, 2000.

[56] M. Elhilali and S. A. Shamma, "A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation," *J. Acoust. Soc. Amer.*, vol. 124, no. 6, pp. 3751–3771, Dec. 2008.

[57] Z. Duan and B. Pardo, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2121–2133, Nov. 2010.

[58] Z. Duan, Y. Zhang, C. Zhang, and Z. Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 766–778, May 2008.

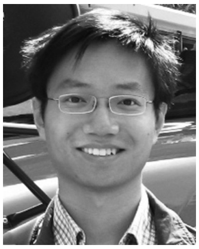
[59] Z. Duan, J. Han, and B. Pardo, "Harmonically informed multi-pitch tracking," in *Proc. 10th Int. Soc. Music Inf. Retrieval*, Kobe, Japan, Oct. 26–30, 2009, pp. 333–338.

[60] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source," in *Proc. 35th Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, Mar. 14–, 2010, pp. 425–428.

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

[62] B. Fox, A. Sabin, B. Pardo, and A. Zopf, "Modeling perceptual similarity of audio signals for blind source separation evaluation," in *Proc. 7th Int. Conf. Ind. Compon. Anal.*, London, U.K., Sep. 9–12, 2007, pp. 454–461.

Zafar Rafii (S'11) is a Ph.D. candidate in electrical engineering and computer science at Northwestern University. He received a Master of Science in electrical engineering, computer science and telecommunications from Ecole Nationale Supérieure de l'Electronique et de ses Applications (ENSEA) in France and a Master of Science in electrical engineering from Illinois Institute of Technology (IIT) in the U.S. He also worked as a research engineer at Audionamix in France and as a research intern at Gracenote in the U.S. His research interests are centered on audio analysis, at the intersection of signal processing, machine learning, and cognitive science.



Zhiyao Duan (S'09–M'13), is an assistant professor in the Electrical and Computer Engineering Department at the University of Rochester. He received his B.S. and M.S. in automation from Tsinghua University, China, in 2004 and 2008, respectively, and his Ph.D. in computer science from Northwestern University in 2013. His research interest is in the broad area of computer audition, i.e., designing computational systems that are capable of analyzing and processing sounds, including music, speech, and environmental sounds. Specific problems that he has been working on include automatic music transcription, multi-pitch analysis, music audio-score alignment, sound source separation, and speech enhancement.

Bryan Pardo (M'07) is an associate professor in the Northwestern University Department of Electrical Engineering and Computer Science. He received a M.Mus. in jazz studies in 2001 and a Ph.D. in computer science in 2005, both from the University of Michigan. He has authored over 50 peer-reviewed publications. He is an associate editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. He has developed speech analysis software for the Speech and Hearing Department of The Ohio State University, statistical software for SPSS, and worked as a machine learning researcher for General Dynamics. While finishing his doctorate, he taught in the Music Department of Madonna University.