# An Overview of Lead and Accompaniment Separation in Music

Zafar Rafii, *Member, IEEE*, Antoine Liutkus, *Member, IEEE*, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, *Student Member, IEEE*, Derry FitzGerald, and Bryan Pardo, *Member, IEEE*

*(Overview Article)*

***Abstract*—Popular music is often composed of an accompaniment and a *lead* component, the latter typically consisting of vocals. Filtering such mixtures to extract one or both components has many applications, such as automatic karaoke and remixing. This particular case of *source separation* yields very specific challenges and opportunities, including the particular complexity of musical structures, but also relevant prior knowledge coming from acoustics, musicology or sound engineering. Due to both its importance in applications and its challenging difficulty, lead and accompaniment separation has been a popular topic in signal processing for decades. In this article, we provide a comprehensive review of this research topic, organizing the different approaches according to whether they are model-based or data-centered. For model-based methods, we organize them according to whether they concentrate on the lead signal, the accompaniment, or both. For data-centered approaches, we discuss the particular difficulty of obtaining data for learning lead separation systems, and then review recent approaches, notably those based on deep learning. Finally, we discuss the delicate problem of evaluating the quality of music separation through adequate metrics and present the results of the largest evaluation, to-date, of lead and accompaniment separation systems. In conjunction with the above, a comprehensive list of references is provided, along with relevant pointers to available implementations and repositories.**

***Index Terms*—Source separation, music, accompaniment, lead, overview.**

## I. Introduction

MUSIC is a major form of artistic expression and plays a central role in the entertainment industry. While digitization and the Internet led to a revolution in the way music reaches its audience [1], [2], there is still much room to improve on how one interacts with musical content, beyond simply controlling the master volume and equalization. The ability to interact with the individual audio objects (e.g., the lead vocals) in a music recording would enable diverse applications such as music upmixing and remixing, automatic karaoke, object-wise equalization, etc.

Most publicly available music recordings (e.g., CDs, YouTube, iTunes, Spotify) are distributed as mono or stereo mixtures with multiple sound objects sharing a track. Therefore, manipulation of individual sound objects requires separation of the stereo audio mixture into several tracks, one for each different sound sources. This process is called *audio source separation* and this overview paper is concerned with an important particular case: isolating the *lead* source—typically, the vocals—from the musical accompaniment (all the rest of the signal).

As a general problem in applied mathematics, source separation has enjoyed tremendous research activity for roughly 50 years and has applications in various fields such as bioinformatics, telecommunications, and audio. Early research focused on so-called *blind* source separation, which typically builds on very weak assumptions about the signals that comprise the mixture in conjunction with very strong assumptions on the way they are mixed. The reader is referred to [3], [4] for a comprehensive review on blind source separation. Typical blind algorithms, e.g., independent component analysis (ICA) [5], [6], depend on assumptions such as: source signals are independent, there are more mixture channels than there are signals, and mixtures are well modeled as a linear combination of signals. While such assumptions are appropriate for some signals like electroencephalograms, they are often violated in audio.

Much research in audio-specific source separation [7], [8] has been motivated by the *speech enhancement* problem [9], which aims to recover clean speech from noisy recordings and can be seen as a particular instance of source separation. In this respect, many algorithms assume the audio background can be modeled as stationary. However, the musical sources are characterized by a very rich, non-stationary spectro-temporal structure. This prohibits the use of such methods. Musical sounds often exhibit highly synchronous evolution over both time and frequency, making overlap in both time and frequency very common. Furthermore, a typical commercial music mix-

Manuscript received July 20, 2017; revised January 4, 2018 and April 3, 2018; accepted April 6, 2018. Date of publication April 12, 2018; date of current version May 8, 2018. This work was supported in part by the Research Programme KAMoulox (ANR-15-CE38-0003-01) funded by ANR and in part by the French State Agency for Research. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tuomas Virtanen. *(Corresponding author: Zafar Rafii.)*

Z. Rafii is with Gracenote, Emeryville, CA 94608 USA (e-mail: zafar.rafii@nielsen.com).

A. Liutkus and F.-R. Stöter are with Inria and LIRMM, University of Montpellier, Montpellier 34090, France (e-mail: antoine.liutkus@inria.fr; fabian-robert.stoeter@audiolabs-erlangen.de).

S. I. Mimilakis is with Fraunhofer IDMT, Ilmenau 98693, Germany (e-mail: mis@idmt.fraunhofer.de).

D. FitzGerald is with the Cork School of Music, Cork Institute of Technology, Cork T12 P928, U.K. (e-mail: Derry.Fitzgerald@cit.ie).

B. Pardo is with Northwestern University, Evanston, IL 60208 USA (e-mail: pardo@northwestern.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.2018.2825440

ture violates all the classical assumptions of ICA. Instruments are correlated (e.g., a chorus of singers), there are more instruments than channels in the mixture, and there are non-linearities in the mixing process (e.g., dynamic range compression). This all has required the development of music-specific algorithms, exploiting available prior information about source structure or mixing parameters [10], [11].

This article provides an overview of nearly 50 years of research on lead and accompaniment separation in music. Due to space constraints and the large variability of the paradigms involved, we cannot delve into detailed mathematical description of each method. Instead, we will convey core ideas and methodologies, grouping approaches according to common features. As with any attempt to impose an *a posteriori* taxonomy on such a large body of research, the resulting classification is arguable. However, we believe it is useful as a roadmap of the relevant literature.

Our objective is not to advocate one methodology over another. While the most recent methods—in particular those based on deep learning—currently show the best performance, we believe that ideas underlying earlier methods may also be inspiring and stimulate new research. This point of view leads us to focus more on the strengths of the methods rather than on their weaknesses.

The rest of the article is organized as follows. In Section II, we present the basic concepts needed to understand the discussion. We then present sections on *model-based* methods that exploit specific knowledge about the lead and/or the accompaniment signals in music to achieve separation. We show in Section III how one body of research is focused on modeling the lead signal as harmonic, exploiting this central assumption for separation. Then, Section IV describes many methods achieving separation using a model that takes the musical accompaniment as *redundant*. In Section V, we show how these two ideas were combined in other studies to achieve separation. Then, we present data-driven approaches in Section VI, which exploit large databases of audio examples where both the isolated lead and accompaniment signals are available. This enables the use of machine learning methods to learn how to separate. In Section VII, we show how the widespread availability of stereo signals may be leveraged to design algorithms that assume centered-panned vocals, but also to improve separation of most methods. Finally, Section VIII is concerned with the problem of how to evaluate the quality of the separation, and provides the results for the largest evaluation campaign to date on this topic.

## II. Fundamental Concepts

We now very briefly describe the basic ideas required to understand this paper, classified into three main categories: signal processing, audio modeling and probability theory. The interested reader is strongly encouraged to delve into the many online courses or textbooks available for a more detailed presentation of these topics, such as [12], [13] for signal processing, [9] for speech modeling, and [14], [15] for probability theory.

### A. Signal Processing

Sound is a series of pressure waves in the air. It is recorded as a *waveform*, a time-series of measurements of the displacement of the microphone diaphragm in response to these pressure waves. Sound is reproduced if a loudspeaker diaphragm is moved according to the recorded waveform. Multichannel signals simply consist of several waveforms, captured by more than one microphone. Typically, music signals are stereophonic, containing two waveforms.

Microphone displacement is typically measured at a fixed *sampling frequency*. In music processing, it is common to have sampling frequencies of 44.1 kHz (the sample frequency on a compact disc) or 48 kHz, which are higher than the typical sampling rates of 16 kHz or 8 kHz used for speech in telephony. This is because musical signals contain much higher frequency content than speech and the goal is aesthetic beauty in addition to basic intelligibility.

A time-frequency (TF) representation of sound is a matrix that encodes the time-varying *spectrum* of the waveform. Its entries are called TF *bins* and encode the varying spectrum of the waveform for all time frames and frequency channels. The most commonly-used TF representation is the short time Fourier transform (STFT) [16], which has complex entries: the angle accounts for the phase, i.e., the actual shift of the corresponding sinusoid at that time bin and frequency bin, and the magnitude accounts for the amplitude of that sinusoid in the signal. The magnitude (or power) of the STFT is called *spectrogram*. When the mixture is multichannel, the TF representation for each channel is computed, leading to a three-dimensional array: frequency, time and channel.

A TF representation is typically used as a first step in processing the audio because sources tend to be less overlapped in the TF representation than in the waveform [17]. This makes it easier to select portions of a mixture that correspond to only a single source. An STFT is typically used because it can be inverted back to the original waveform. Therefore, modifications made to the STFT can be used to create a modified waveform. Generally, a linear mixing process is considered, i.e., the mixture signal is equal to the sum of the source signals. Since the Fourier transform is a linear operation, this equality holds for the STFT. While that is not the case for the magnitude (or power) of the STFT, it is commonly assumed that the spectrograms of the sources sum to the spectrogram of the mixture.

In many methods, the separated sources are obtained by *filtering* the mixture. This can be understood as performing some equalization on the mixture, where each frequency is attenuated or kept intact. Since both the lead and the accompaniment signals change over time, the filter also changes. This is typically done using a TF *mask*, which, in its simplest form, is defined as the gain between 0 and 1 to apply on each element of the TF representation of the mixture (e.g., an STFT) in order to estimate the desired signal. Loosely speaking, it can be understood as an equalizer whose setting changes every few milliseconds. After multiplication of the mixture by a mask, the separated signal is recovered through an inverse TF transform. In the multichannel setting, more sophisticated filters may be designed that

incorporate some delay and combine different channels; this is usually called *beamforming*. In the frequency domain, this is often equivalent to using complex matrices to multiply the mixture TF representation with, instead of just scalars between 0 and 1.

In practice, masks can be designed to filter the mixture in several ways. One may estimate the spectrogram for a single source or component, e.g., the accompaniment, and subtract it from the mixture spectrogram, e.g., in order to estimate the lead [18]. Another way would be to estimate separate spectrograms for both lead and accompaniment and combine them to yield a mask. For instance, a TF mask for the lead can be taken as the proportion of the lead spectrogram over the sum of both spectrograms, at each TF bin. Such filters are often called *Wiener filters* [19] or *ratio masks*. How they are calculated may involve some additional techniques like exponentiation and may be understood according to assumptions regarding the underlying statistics of the sources. For recent work in this area, and many useful pointers in designing such masks, the reader is referred to [20].

### B. Audio and Speech Modeling

It is typical in audio processing to describe audio waveforms as belonging to one of two different categories, which are *sinusoidal signals*—or pure tones—and *noise*. Actually, both are just the two extremes in a continuum of varying *predictability*: on the one hand, the shape of a sinusoidal wave in the future can reliably be guessed from previous samples. On the other hand, white noise is *defined* as an unpredictable signal and its spectrogram has constant energy everywhere. Different noise profiles may then be obtained by attenuating the energy of some frequency regions. This in turn induces some predictability in the signal, and in the extreme case where all the energy content is concentrated in one frequency, a pure tone is obtained.

A waveform may always be modeled as some *filter* applied on some *excitation signal*. Usually, the filter is assumed to vary smoothly across frequencies, hence modifying only what is called *the spectral envelope* of the signal, while the excitation signal comprises the rest. This is the basis for the *source-filter* model [21], which is of great importance in speech modeling, and thus also in vocal separation. As for speech, the filter is created by the shape of the vocal tract. The excitation signal is made of the glottal pulses generated by the vibration of the vocal folds. This results into *voiced* speech sounds made of time-varying harmonic/sinusoidal components. The excitation signal can also be the air flow passing through some constriction of the vocal tract. This results into *unvoiced*, noise-like, speech sounds. In this context, vowels are said to be voiced and tend to feature many sinusoids, while some phonemes such as fricatives are unvoiced and noisier.

A classical tool for dissociating the envelope from the excitation is the *cepstrum* [22]. It has applications for estimating the fundamental frequency [23], [24], for deriving the Mel-frequency cepstral coefficients (MFCC) [25], or for filtering signals through a so-called *liftering* operation [26] that enables modifications of either the excitation or the envelope parts through the source-filter paradigm.

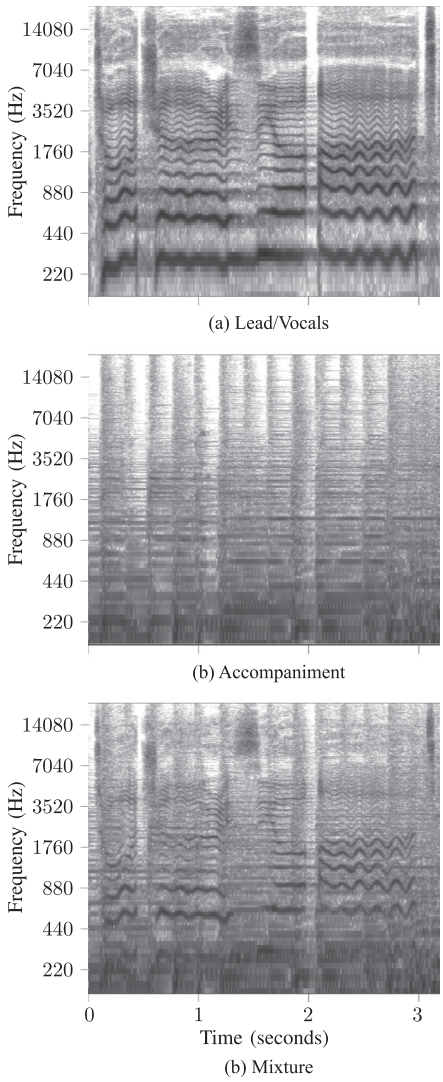



Fig. 1. Examples of spectrograms from an excerpt of the track "The Wrong'Uns - Rothko" from MUSDB18 dataset. The two sources to be separated are depicted in (a) and (b), and its mixture in (c). The vocals (a) are mostly harmonic and often well described by a source-filter model in which an excitation signal is filtered according to the vocal tract configuration. The accompaniment signal (b) features more diversity, but usually does not feature as much vibrato as for the vocals, and most importantly is seen to be *denser* and also *more redundant*. All spectrograms have log-compressed amplitudes as well as log-scaled frequency axis.

An advantage of the source-filter model approach is indeed that one can dissociate the pitched content of the signal, embodied by the position of its harmonics, from its TF envelope which describes where the energy of the sound lies. In the case of vocals, it yields the ability to distinguish between the actual note being sung (pitch content) and the phoneme being uttered (mouth and vocal tract configuration), respectively. One key feature of vocals is they typically exhibit great variability in fundamental frequency over time. They can also exhibit larger *vibratos* (fundamental frequency modulations) and *tremolos* (amplitude modulations) in comparison to other instruments, as seen in the top spectrogram in Fig. 1.

A particularity of musical signals is that they typically consist of sequences of pitched notes. A sound gives the perception of having a pitch if the majority of the energy in the audio signal is

at frequencies located at integer multiples of some fundamental frequency. These integer multiples are called *harmonics*. When the fundamental frequency changes, the frequencies of these harmonics also change, yielding the typical comb spectrograms of harmonic signals, as depicted in the top spectrogram in Fig. 1. Another noteworthy feature of sung melodies over simple speech is that their fundamental frequencies are, in general, located at precise frequency values corresponding to the musical key of the song. These very peculiar features are often exploited in separation methods. For simplicity reasons, we use the terms *pitch* and *fundamental frequency* interchangeably throughout the paper.

### C. Probability Theory

Probability theory [14], [27] is an important framework for designing many data analysis and processing methods. Many of the methods described in this article use it and it is far beyond the scope of this paper to present it rigorously. For our purpose, it will suffice to say that the *observations* consist of the mixture signals. On the other hand, the *parameters* are any relevant feature about the source signal (such as pitch or time-varying envelope) or how the signals are mixed (e.g., the panning position). These parameters can be used to derive estimates about the target lead and accompaniment signals.

We understand a probabilistic *model* as a function of both the observations and the parameters: it describes how likely the observations are, given the parameters. For instance, a flat spectrum is likely under the noise model, and a mixture of comb spectrograms is likely under a harmonic model with the appropriate pitch parameters for the sources. When the observations are given, variation in the model depends only on the parameters. For some parameter value, it tells how likely the observations are. Under a harmonic model for instance, pitch may be estimated by finding the pitch parameter that makes the observed waveform as likely as possible. Alternatively, we may want to choose between several possible models such as voiced or unvoiced. In such cases, *model selection* methods are available, such as the Bayesian information criterion (BIC) [28].

Given these basic ideas, we briefly mention two models that are of particular importance. Firstly, the hidden Markov model (HMM) [15], [29] is relevant for time-varying observations. It basically defines several *states*, each one related to a specific model and with some probabilities for transitions between them. For instance, we could define as many states as possible notes played by the lead guitar, each one associated with a typical spectrum. The *Viterbi algorithm* is a dynamic programming method which actually estimates the most likely sequence of states given a sequence of observations [30]. Secondly, the Gaussian mixture model (GMM) [31] is a way to approximate any distribution as a weighted sum of Gaussians. It is widely used in clustering, because it works well with the celebrated Expectation-Maximization (EM) algorithm [32] to assign one particular cluster to each data point, while automatically estimating the clusters parameters. As we will see later, many methods work by assigning each TF bin to a given source in a similar way.

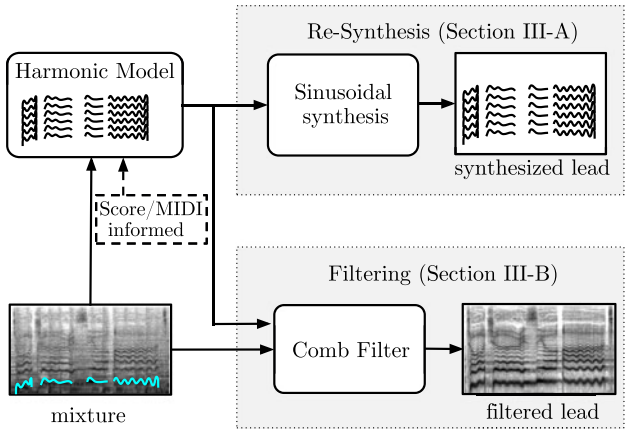



Fig. 2. The approaches based on a *harmonic assumption* for vocals. In a first analysis step, the fundamental frequency of the lead signal is extracted. From it, a separation is obtained either by resynthesis (Section III-A), or by filtering the mixture (Section III-B).

## III. Modeling the Lead Signal: Harmonicity

As mentioned in Section II-B, one particularity of vocals is their production by the vibration of the vocal folds, further filtered by the vocal tract. As a consequence, sung melodies are *mostly* harmonic, as depicted in Fig. 1, and therefore have a fundamental frequency. If one can track the pitch of the vocals, one can then estimate the energy at the harmonics of the fundamental frequency and reconstruct the voice. This is the basis of the oldest methods (as well as some more recent methods) we are aware of for separating the lead signal from a musical mixture.

Such methods are summarized in Fig. 2. In a first step, the objective is to get estimates of the time-varying fundamental frequency for the lead at each time frame. A second step in this respect is then to track this fundamental frequency over time, in other words, to find the best sequence of estimates, in order to identify the melody line. This can done either by a suitable pitch detection method, or by exploiting the availability of the score. Such algorithms typically assume that the lead corresponds to the harmonic signal with strongest amplitude. For a review on the particular topic of melody extraction, the reader is referred to [33].

From this starting point, we can distinguish between two kinds of approaches, depending on how they exploit the pitch information.

### A. Analysis-Synthesis Approaches

The first option to obtain the separated lead signal is to resynthesize it using a sinusoidal model. A sinusoidal model decomposes the sound with a set of sine waves of varying frequency and amplitude. If one knows the fundamental frequency of a pitched sound (like a singing voice), as well as the spectral envelope of the recording, then one can reconstruct the sound by making a set of sine waves whose frequencies are those of the harmonics of the fundamental frequency, and whose amplitudes are estimated from the spectral envelope of the audio. While the spectral envelope of the recording is generally not exactly the same as the spectral envelope of the target source, it can

be a reasonable approximation, especially assuming that different sources do not overlap too much with each other in the TF representation of the mixture.

This idea allows for time-domain processing and was used in the earliest methods we are aware of. In 1973, Miller proposed in [34] to use the homomorphic vocoder [35] to separate the excitation function and impulse response of the vocal tract. Further refinements include segmenting parts of the signal as voiced, unvoiced, or silences using a heuristic program and manual interaction. Finally, cepstral liftering [26] was exploited to compensate for the noise or accompaniment.

Similarly, Maher used an analysis-synthesis approach in [36], assuming the mixtures are composed of only two harmonic sources. In his case, pitch detection was performed on the STFT and included heuristics to account for possibly colliding harmonics. He finally resynthesized each musical voice with a sinusoidal model.

Wang proposed instantaneous and frequency-warped techniques for signal parameterization and source separation, with application to voice separation in music [37], [38]. He introduced a frequency-locked loop algorithm which uses multiple harmonically constrained trackers. He computed the estimated fundamental frequency from a maximum-likelihood weighting of the tracking estimates. He was then able to estimate harmonic signals such as voices from complex mixtures.

Meron and Hirose proposed to separate singing voice and piano accompaniment [39]. In their case, prior knowledge consisting of musical scores was considered. Sinusoidal modeling as described in [40] was used.

Ben-Shalom and Dubnov proposed to filter an instrument or a singing voice out in such a way [41]. They first used a score alignment algorithm [42], assuming a known score. Then, they used the estimated pitch information to design a filter based on a harmonic model [43] and performed the filtering using the linear constraint minimum variance approach [44]. They additionally used a heuristic to deal with the unvoiced parts of the singing voice.

Zhang and Zhang proposed an approach based on harmonic structure modeling [45], [46]. They first extracted harmonic structures for singing voice and background music signals using a sinusoidal model [43], by extending the pitch estimation algorithm in [47]. Then, they used the clustering algorithm in [48] to learn harmonic structure models for the background music signals. Finally, they extracted the harmonic structures for all the instruments to reconstruct the background music signals and subtract them from the mixture, leaving only the singing voice signal.

More recently, Fujihara *et al.* proposed an accompaniment reduction method for singer identification [49], [50]. After fundamental frequency estimation using [51], they extracted the harmonic structure of the melody, i.e., the power and phase of the sinusoidal components at fundamental frequency and harmonics. Finally, they resynthesized the audio signal of the melody using the sinusoidal model in [52].

Similarly, Mesaros *et al.* proposed a vocal separation method to help with singer identification [53]. They first applied a melody transcription system [54] which estimates the melody line with the corresponding MIDI note numbers. Then, they performed sinusoidal resynthesis, estimating amplitudes and phases from the polyphonic signal.

In a similar manner, Duan *et al.* proposed to separate harmonic sources, including singing voices, by using harmonic structure models [55]. They first defined an average harmonic structure model for an instrument. Then, they learned a model for each source by detecting the spectral peaks using a cross-correlation method [56] and quadratic interpolation [57]. Then, they extracted the harmonic structures using BIC and a clustering algorithm [48]. Finally, they separated the sources by re-estimating the fundamental frequencies, re-extracting the harmonics, and reconstructing the signals using a phase generation method [58].

Lagrange *et al.* proposed to formulate lead separation as a graph partition problem [59], [60]. They first identified peaks in the spectrogram and grouped the peaks into clusters by using a similarity measure which accounts for harmonically related peaks, and the normalized cut criterion [61] which is used for segmenting graphs in computer vision. They finally selected the cluster of peaks which corresponds to a predominant harmonic source and resynthesized it using a bank of sinusoidal oscillators.

Ryynänen *et al.* proposed to separate accompaniment from polyphonic music using melody transcription for karaoke applications [62]. They first transcribed the melody into a MIDI note sequence and a fundamental frequency trajectory, using the method in [63], an improved version of the earlier method [54]. Then, they used sinusoidal modeling to estimate, resynthesize, and remove the lead vocals from the musical mixture, using the quadratic polynomial-phase model in [64].

### *B. Comb-Filtering Approaches*

Using sinusoidal synthesis to generate the lead signal suffers from a typical *metallic* sound quality, which is mostly due to discrepancies between the estimated excitation signals of the lead signal compared to the ground truth. To address this issue, an alternative approach is to exploit harmonicity in another way, by filtering out everything from the mixture that is not located close to the detected harmonics.

Li and Wang proposed to use a vocal/non-vocal classifier and a predominant pitch detection algorithm [65], [66]. They first detected the singing voice by using a spectral change detector [67] to partition the mixture into homogeneous portions, and GMMs on MFCCs to classify the portions as vocal or non-vocal. Then, they used the predominant pitch detection algorithm in [68] to detect the pitch contours from the vocal portions, extending the multi-pitch tracking algorithm in [69]. Finally, they extracted the singing voice by decomposing the vocal portions into TF units and labeling them as singing or accompaniment dominant, extending the speech separation algorithm in [70].

Han and Raphael proposed an approach for desoloing a recording of a soloist with an accompaniment given a musical score and its time alignment with the recording [71]. They derived a mask [72] to remove the solo part after using an EM algorithm to estimate its melody, that exploits the score as side information.

Hsu *et al.* proposed an approach which also identifies and separates the unvoiced singing voice [73], [74]. Instead of pro-

cessing in the STFT domain, they use the perceptually motivated gammatone filter-bank as in [66], [70]. They first detected accompaniment, unvoiced, and voiced segments using an HMM and identified voice-dominant TF units in the voiced frames by using the singing voice separation method in [66], using the predominant pitch detection algorithm in [75]. Unvoiced-dominant TF units were identified using a GMM classifier with MFCC features learned from training data. Finally, filtering was achieved with spectral subtraction [76].

Raphael and Han then proposed a classifier-based approach to separate a soloist from accompanying instruments using a time-aligned symbolic musical score [77]. They built a tree-structured classifier [78] learned from labeled training data to classify TF points in the STFT as belonging to solo or accompaniment. They additionally constrained their classifier to estimate masks having a connected structure.

Cano *et al.* proposed various approaches for solo and accompaniment separation. In [79], they separated saxophone melodies from mixtures with piano and/or orchestra by using a melody line detection algorithm, incorporating information about typical saxophone melody lines. In [80]–[82], they proposed to use the pitch detection algorithm in [83]. Then, they refined the fundamental frequency and the harmonics, and created a binary mask for the solo and accompaniment. They finally used a post-processing stage to refine the separation. In [84], they included a noise spectrum in the harmonic refinement stage to also capture noise-like sounds in vocals. In [85], they additionally included common amplitude modulation characteristics in the separation scheme.

Bosch *et al.* proposed to separate the lead instrument using a musical score [86]. After a preliminary alignment of the score to the mixture, they estimated a score confidence measure to deal with local misalignments and used it to guide the predominant pitch tracking. Finally, they performed low-latency separation based on the method in [87], by combining harmonic masks derived from the estimated pitch and additionally exploiting stereo information as presented later in Section VII.

Vaneph *et al.* proposed a framework for vocal isolation to help spectral editing [88]. They first used a voice activity detection process based on a deep learning technique [89]. Then, they used pitch tracking to detect the melodic line of the vocal and used it to separate the vocal and background, allowing a user to provide manual annotations when necessary.

### C. Shortcomings

As can be seen, explicitly assuming that the lead signal is harmonic led to an important body of research. While the aforementioned methods show excellent performance when their assumptions are valid, their performance can drop significantly in adverse, but common situations.

Firstly, vocals are not always purely harmonic as they contain unvoiced phonemes that are not harmonic. As seen above, some methods already handle this situation. However, vocals can also be whispered or saturated, both of which are difficult to handle with a harmonic model.

Secondly, methods based on the harmonic model depend on the quality of the pitch detection method. If the pitch detector switches from following the pitch of the lead (e.g., the voice) to another instrument, the wrong sound will be isolated from the mix. Often, pitch detectors assume the lead signal is the *loudest* harmonic sound in the mix. Unfortunately, this is not always the case. Another instrument may be louder or the lead may be silent for a passage. The tendency to follow the pitch of the wrong instrument can be mitigated by applying constraints on the pitch range to estimate and by using a perceptually relevant weighting filter before performing pitch tracking. Of course, these approaches do not help when the lead signal is silent.

## IV. Modeling the Accompaniment: Redundancy

In the previous section, we presented methods whose main focus was the modeling of a harmonic lead melody. Most of these studies did not make modeling the accompaniment a core focus. On the contrary, it was often dealt with as adverse noise to which the harmonic processing method should be robust to.

In this section, we present another line of research which concentrates on modeling the accompaniment under the assumption it is somehow more *redundant* than the lead signal. This assumption stems from the fact that musical accompaniments are often highly structured, with elements being repeated many times. Such repetitions can occur at the note level, in terms of rhythmic structure, or even from a harmonic point of view: instrumental notes are often constrained to have their pitch lie in a small set of frequencies. Therefore, modeling and removing the redundant elements of the signal are assumed to result in removal of the accompaniment.

In this paper, we identify three families of methods that exploit the redundancy of the accompaniment for separation.

### A. Grouping Low-Rank Components

The first set of approaches we consider is the identification of redundancy in the accompaniment through the assumption that its spectrogram may be well represented by only a few components. Techniques exploiting this idea then focus on algebraic methods that decompose the mixture spectrogram into the product of a few template spectra activated over time. One way to do so is via non-negative matrix factorization (NMF) [90], [91], which incorporates non-negative constraints. In Fig. 3, we picture methods exploiting such techniques. After factorization, we obtain several spectra, along with their activations over time. A subsequent step is the clustering of these spectra (and activations) into the lead or the accompaniment. Separation is finally performed by deriving Wiener filters to estimate the lead and the accompaniment from the mixture. For related applications of NMF in music analysis, the reader is referred to [92]–[94].

Vembu and Baumann proposed to use NMF (and also ICA [95]) to separate vocals from mixtures [96]. They first discriminated between vocal and non-vocal sections in a mixture by using different combinations of features, such as MFCCs [25], perceptual linear predictive (PLP) coefficients [97], and log frequency power coefficients (LFPC) [98], and training two classifiers, namely neural networks and support vector machines

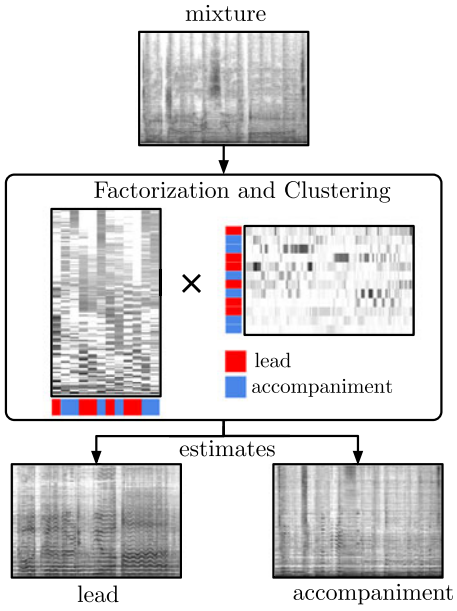



Fig. 3. The approaches based on a *low-rank* assumption. Non-negative matrix factorization (NMF) is used to identify *components* from the mixture, that are subsequently clustered into lead or accompaniment. Additional constraints may be incorporated.

(SVM). They then applied redundancy reduction techniques on the TF representation of the mixture to separate the sources [99], by using NMF (or ICA). The components were then grouped as vocal and non-vocal by reusing a vocal/non-vocal classifier with MFCC, LFPC, and PLP coefficients.

Chanrungutai and Ratanamahatana proposed to use NMF with automatic component selection [100], [101]. They first decomposed the mixture spectrogram using NMF with a fixed number of basis components. They then removed the components with brief rhythmic and long-lasting continuous events, assuming that they correspond to instrumental sounds. They finally used the remaining components to reconstruct the singing voice, after refining them using a high-pass filter.

Marxer and Janer proposed an approach based on a Tikhonov regularization [102] as an alternative to NMF, for singing voice separation [103]. Their method sacrificed the non-negativity constraints of the NMF in exchange for a computationally less expensive solution for spectrum decomposition, making it more interesting in low-latency scenarios.

Yang *et al.* proposed a Bayesian NMF approach [104], [105]. Following the approaches in [106] and [107], they used a Poisson distribution for the likelihood function and exponential distributions for the model parameters in the NMF algorithm, and derived a variational Bayesian EM algorithm [32] to solve the NMF problem. They also adaptively determined the number of bases from the mixture. They finally grouped the bases into singing voice and background music by using a *k*-means clustering algorithm [108] or an NMF-based clustering algorithm.

In a different manner, Smaragdis and Mysore proposed a user-guided approach for removing sounds from mixtures by humming the target sound to be removed, for example a vocal track [109]. They modeled the mixture using probabilistic latent component analysis (PLCA) [110], another equivalent

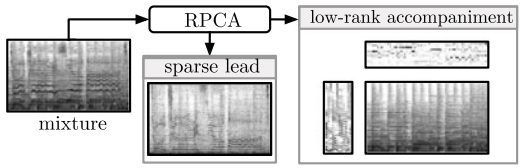



Fig. 4. The approaches based on a *low-rank accompaniment, sparse vocals* assumption. As opposed to methods based on NMF, methods based on robust principal component analysis (RPCA) assume the lead signal has a sparse and non-structured spectrogram.

formulation of NMF. One key feature of exploiting user input was to facilitate the grouping of components into vocals and accompaniment, as humming helped to identify some of the parameters for modeling the vocals.

Nakamuray and Kameoka proposed an $L_p$-norm NMF [111], with $p$ controlling the sparsity of the error. They developed an algorithm for solving this NMF problem based on the auxiliary function principle [112], [113]. Setting an adequate number of bases and $p$ taken as small enough allowed them to estimate the accompaniment as the low-rank decomposition, and the singing voice as the error of the approximation, respectively. Note that, in this case, the singing voice was not explicitly modeled as a sparse component but rather corresponded to the error which happened to be constrained as sparse. The next subsection will actually deal with approaches that explicitly model the vocals as the sparse component.

### B. Low-Rank Accompaniment, Sparse Vocals

The methods presented in the previous section first compute a decomposition of the mixture into many components that are sorted *a posteriori* as accompaniment or lead. As can be seen, this means they make a low-rank assumption for the accompaniment, but typically *also for the vocals*. However, as can for instance be seen on Fig. 1, the spectrogram for the vocals do exhibit much more freedom than accompaniment, and experience shows they are not adequately described by a small number of spectral bases. For this reason, another track of research depicted in Fig. 4 focused on using a low-rank assumption on the accompaniment *only*, while assuming the vocals are *sparse and not structured*. This loose assumption means that only a few coefficients from their spectrogram should have significant magnitude, and that they should not feature significant redundancy. Those ideas are in line with robust principal component analysis (RPCA) [114], which is the mathematical tool used by this body of methods, initiated by Huang *et al.* for singing voice separation [115]. It decomposes a matrix into a sparse and low-rank component.

Sprechmann *et al.* proposed an approach based on RPCA for online singing voice separation [116]. They used ideas from convex optimization [117], [118] and multi-layer neural networks [119]. They presented two extensions of RPCA and robust NMF models [120]. They then used these extensions in a multi-layer neural network framework which, after an initial training stage, allows online source separation.

Jeong and Lee proposed two extensions of the RPCA model to improve the estimation of vocals and accompaniment from the sparse and low-rank components [121]. Their first extension

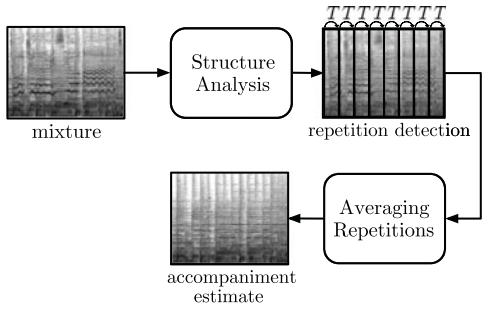



Fig. 5. The approaches based on a *repetition* assumption for accompaniment. In a first analysis step, repetitions are identified. Then, they are used to build an estimate for the accompaniment spectrogram and proceed to separation.

included the Schatten $p$ and $\ell_p$ norms as generalized nuclear norm optimizations [122]. They also suggested a pre-processing stage based on logarithmic scaling of the mixture TF representation to enhance the RPCA.

Yang also proposed an approach based on RPCA with dictionary learning for recovering low-rank components [123]. He introduced a multiple low-rank representation following the observation that elements of the singing voice can also be recovered by the low-rank component. He first incorporated online dictionary learning methods [124] in his methodology to obtain prior information about the structure of the sources and then incorporated them into the RPCA model.

Chan and Yang then extended RPCA to complex and quaternionic cases with application to singing voice separation [125]. They extended the principal component pursuit (PCP) [114] for solving the RPCA problem by presenting complex and quaternionic proximity operators for the $\ell_1$ and trace-norm regularizations to account for the missing phase information.

### C. Repetitions Within the Accompaniment

While the rationale behind low-rank methods for lead-accompaniment separation is to exploit the idea that the musical background should be redundant, adopting a low-rank model is not the only way to do it. An alternate way to proceed is to exploit the musical *structure* of songs, to find *repetitions* that can be utilized to perform separation. Just like in RPCA-based methods, the accompaniment is then assumed to be the only source for which repetitions will be found. The unique feature of the methods described here is they combine music structure analysis [126]–[128] with particular ways to exploit the identification of repeated parts of the accompaniment.

Rafii *et al.* proposed the REpeating Pattern Extraction Technique (REPET) to separate the accompaniment by assuming it is repeating [129]–[131], which is often the case in popular music. This approach, which is representative of this line of research, is represented on Fig. 5. First, a repeating period is extracted by a music information retrieval system, such as a beat spectrum [132] in this case. Then, this extracted information is used to estimate the spectrogram of the accompaniment through an averaging of the identified repetitions. From this, a filter is derived.

Seetharaman *et al.* [133] leveraged the two dimensional Fourier transform (2DFT) of the spectrogram to create an algorithm very similar to REPET. The properties of the 2DFT let them separate the periodic background from the non-periodic vocal melody by deleting peaks in the 2DFT. This eliminated the need to create an explicit model of the periodic audio and without the need to find the period of repetition, both of which are required in REPET.

Liutkus *et al.* adapted the REPET approach in [129], [130] to handle repeating structures varying along time by modeling the repeating patterns only locally [131], [134]. They first identified a repeating period for every time frame by computing a beat spectrogram as in [132]. Then they estimated the spectrogram of the accompaniment by averaging the time frames in the mixture spectrogram at their local period rate, for every TF bin. From this, they finally extracted the repeating structure by deriving a TF mask.

Rafii *et al.* further extended the REPET approaches in [129], [130] and [134] to handle repeating structures that are not periodic. To do this, they proposed the REPET-SIM method in [131], [135] to identify repeating frames for every time frame by computing a self-similarity matrix, as in [136]. Then, they estimated the accompaniment spectrogram at every TF bin by averaging the neighbors identified thanks to that similarity matrix. An extension for real-time processing was presented in [137] and a version exploiting user interaction was proposed in [138]. A method close to REPET-SIM was also proposed by FitzGerald in [139].

Liutkus *et al.* proposed the Kernel Additive modeling (KAM) [140], [141] as a framework which generalizes the REPET approaches in [129]–[131], [134], [135]. They assumed that a source at a TF location can be modeled using its values at other locations through a specified kernel which can account for features such as periodicity, self-similarity, stability over time or frequency, etc. This notably enabled modeling of the accompaniment using more than one repeating pattern. Liutkus *et al.* also proposed a light version using a fast compression algorithm to make the approach more scalable [142]. The approach was also used for interference reduction in music recordings [143], [144].

With the same idea of exploiting intra-song redundancies for singing voice separation, but through a very different methodology, Moussallam *et al.* assumed in [145] that all the sources can be decomposed sparsely in the same dictionary and used a matching pursuit greedy algorithm [146] to solve the problem. They integrated the separation process in the algorithm by modifying the atom selection criterion and adding a decision to assign a chosen atom to the repeated source or to the lead signal.

Deif *et al.* proposed to use multiple median filters to separate vocals from music recordings [147]. They augmented the approach in [148] with diagonal median filters to improve the separation of the vocal component. They also investigated different filter lengths to further improve the separation.

Lee *et al.* also proposed to use the KAM approach [149]–[152]. They applied the $\beta$-order minimum mean square error (MMSE) estimation [153] to the back-fitting algorithm in KAM to improve the separation. They adaptively calculated a perceptually weighting factor $\alpha$ and the singular value decomposition (SVD)-based factorized spectral amplitude exponent $\beta$ for each kernel component.

### D. Shortcomings

While methods focusing on harmonic models for the lead often fall short in their expressive power for the accompaniment, the methods we reviewed in this section are often observed to suffer exactly from the converse weakness, namely they do not provide an adequate model for the lead signal. Hence, the separated vocals often will feature interference from unpredictable parts from the accompaniment, such as some percussion or effects which occur infrequently.

Furthermore, even if the musical accompaniment will exhibit more redundancy, the vocals part will also be redundant to some extent, which is poorly handled by these methods. When the lead signal is not vocals but played by some lead instrument, its redundancy is even more pronounced, because the notes it plays lie in a reduced set of fundamental frequencies. Consequently, such methods would include the redundant parts of the lead within the accompaniment estimate, for example, a steady humming by a vocalist.

## V. Joint Models for Lead and Accompaniment

In the previous sections, we reviewed two important bodies of literature, focused on modeling either the lead or the accompaniment parts of music recordings, respectively. While each approach showed its own advantages, it also featured its own drawbacks. For this reason, some researchers devised methods combining ideas for modeling both the lead and the accompaniment sources, and thus benefiting from both approaches. We now review this line of research.

### A. Using Music Structure Analysis to Drive Learning

The first idea we find in the literature is to augment methods for accompaniment modeling with the prior identification of sections where the vocals are present or absent. In the case of the low rank models discussed in Sections IV-A and IV-B, such a strategy indeed dramatically improves performance.

Raj *et al.* proposed an approach in [154] that is based on the PLCA formulation of NMF [155], and extends their prior work [156]. The parameters for the frequency distribution of the background music are estimated from the background music-only segments, and the rest of the parameters from the singing voice+background music segments, assuming a priori identified vocal regions.

Han and Chen also proposed a similar approach for melody extraction based on PLCA [157], which includes a further estimate of the melody from the vocals signal by an autocorrelation technique similar to [158].

Gómez *et al.* proposed to separate the singing voice from the guitar accompaniment in flamenco music to help with melody transcription [159]. They first manually segmented the mixture into vocal and non-vocal regions. They then learned percussive and harmonic bases from the non-vocal regions by using an unsupervised NMF percussive/harmonic separation approach [93], [160]. The vocal spectrogram was estimated by keeping the learned percussive and harmonic bases fixed.

Papadopoulos and Ellis proposed a signal-adaptive formulation of RPCA which incorporates music content information

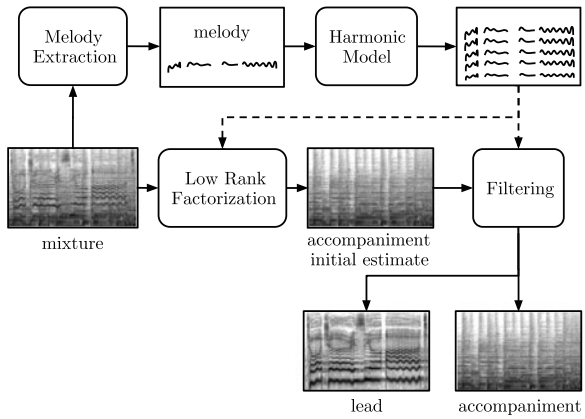



Fig. 6. Factorization informed with the melody. First, melody extraction is performed on the mixture. Then, this information is used to drive the estimation of the accompaniment: TF bins pertaining to the lead should not be taken into account for estimating the accompaniment model.

to guide the recovery of the sparse and low-rank components [161]. Prior musical knowledge, such as predominant melody, is used to regularize the selection of active coefficients during the optimization procedure.

In a similar manner, Chan *et al.* proposed to use RPCA with vocal activity information [162]. They modified the RPCA algorithm to constraint parts of the input spectrogram to be non-sparse to account for the non-vocal parts of the singing voice.

A related method was proposed by Jeong and Lee in [163], using RPCA with a weighted $l_1$-norm. They replaced the uniform weighting between the low-rank and sparse components in the RPCA algorithm by an adaptive weighting based on the variance ratio between the singing voice and the accompaniment. One key element of the method is to incorporate vocal activation information in the weighting.

### B. Factorization With a Known Melody

While using only the knowledge of vocal activity as described above already yields an increase of performance over methods operating blindly, many authors went further to also incorporate the fact that vocals often have a strong melody line. Some redundant model is then assumed for the accompaniment, while also enforcing a harmonic model for the vocals.

An early method to achieve this is depicted in Fig. 6 and was proposed by Virtanen *et al.* in [164]. They estimated the pitch of the vocals in the mixture by using a melody transcription algorithm [63] and derived a binary TF mask to identify where vocals are not present. They then applied NMF on the remaining non-vocal segments to learn a model for the background.

Wang and Ou also proposed an approach which combines melody extraction and NMF-based soft masking [165]. They identified accompaniment, unvoiced, and voiced segments in the mixture using an HMM model with MFCCs and GMMs. They then estimated the pitch of the vocals from the voiced segments using the method in [166] and an HMM with the Viterbi algorithm as in [167]. They finally applied a soft mask to separate voice and accompaniment.

Rafii *et al.* investigated the combination of an approach for modeling the background and an approach for modeling the

melody [168]. They modeled the background by deriving a rhythmic mask using the REPET-SIM algorithm [135] and the melody by deriving a harmonic mask using a pitch-based algorithm [169]. They proposed a parallel and a sequential combination of those algorithms.

Venkataramani *et al.* proposed an approach combining sinusoidal modeling and matrix decomposition, which incorporates prior knowledge about singer and phoneme identity [170]. They applied a predominant pitch algorithm on annotated sung regions [171] and performed harmonic sinusoidal modeling [172]. Then, they estimated the spectral envelope of the vocal component from the spectral envelope of the mixture using a phoneme dictionary. After that, a spectral envelope dictionary representing sung vowels from song segments of a given singer was learned using an extension of NMF [173], [174]. They finally estimated a soft mask using the singer-vowel dictionary to refine and extract the vocal component.

Ikemiya *et al.* proposed to combine RPCA with pitch estimation [175], [176]. They derived a mask using RPCA [115] to separate the mixture spectrogram into singing voice and accompaniment components. They then estimated the fundamental frequency contour from the singing voice component based on [177] and derived a harmonic mask. They integrated the two masks and resynthesized the singing voice and accompaniment signals. Dobashi *et al.* then proposed to use that singing voice separation approach in a music performance assistance system [178].

Hu and Liu proposed to combine approaches based on matrix decomposition and pitch information for singer identification [179]. They used non-negative matrix partial co-factorization [173], [180] which integrates prior knowledge about the singing voice and the accompaniment, to separate the mixture into singing voice and accompaniment portions. They then identified the singing pitch from the singing voice portions using [181] and derived a harmonic mask as in [182], and finally reconstructed the singing voice using a missing feature method [183]. They also proposed to add temporal and sparsity criteria to their algorithm [184].

That methodology was also adopted by Zhang *et al.* in [185], that followed the framework of the pitch-based approach in [66], by performing singing voice detection using an HMM classifier, singing pitch detection using the algorithm in [186], and singing voice separation using a binary mask. Additionally, they augmented that approach by analyzing the latent components of the TF matrix using NMF in order to refine the singing voice and accompaniment.

Zhu *et al.* [187] proposed an approach which is also representative of this body of literature, with the pitch detection algorithm being the one in [181] and binary TF masks used for separation after NMF.

### C. Joint Factorization and Melody Estimation

The methods presented above put together the ideas of modeling the lead (typically the vocals) as featuring a melodic harmonic line and the accompaniment as redundant. As such, they already exhibit significant improvement over approaches only applying one of these ideas as presented in Sections III and IV, respectively. However, these methods above are still restricted in the sense that the analysis performed on each side cannot help improve the other one. In other words, the estimation of the models for the lead and the accompaniment are done sequentially. Another idea is to proceed *jointly*.

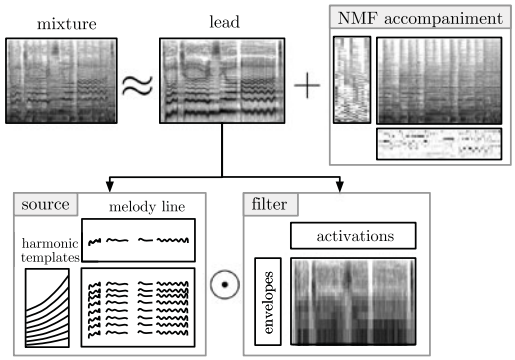



Fig. 7. Joint estimation of the lead and accompaniment, the former one as a source-filter model and the latter one as an NMF model.

A seminal work in this respect was done by Durrieu *et al.* using a source-filter and NMF model [188]–[190], depicted in Fig. 7. Its core idea is to decompose the mixture spectrogram as the sum of two terms. The first term accounts for the lead and is inspired by the source-filter model described in Section II: it is the element-wise product of an *excitation* spectrogram with a *filter* spectrogram. The former one can be understood as harmonic combs activated by the melodic line, while the latter one modulates the envelope and is assumed low-rank because few phonemes are used. The second term accounts for the accompaniment and is modeled with a standard NMF. In [188]–[190], they modeled the lead by using a GMM-based model [191] and a glottal source model [192], and the accompaniment by using an instantaneous mixture model [193] leading to an NMF problem [94]. They jointly estimated the parameters of their models by maximum likelihood estimation using an iterative algorithm inspired by [194] with multiplicative update rules developed in [91]. They also extracted the melody by using an algorithm comparable to the Viterbi algorithm, before re-estimating the parameters and finally performing source separation using Wiener filters [195]. In [196], they proposed to adapt their model for user-guided source separation.

The joint modeling of the lead and accompaniment parts of a music signal was also considered by Fuentes *et al.* in [197], that introduced the idea of using a log-frequency TF representation called the constant-Q transform (CQT) [198]–[200]. The advantage of such a representation is that a change in pitch corresponds to a simple translation in the TF plane, instead of a scaling as in the STFT. This idea was used along the creation of a user interface to guide the decomposition, in line with what was done in [196].

Joder and Schuller used the source-filter NMF model in [201], additionally exploiting MIDI scores [202]. They synchronized the MIDI scores to the audio using the alignment algorithm in [203]. They proposed to exploit the score information through two types of constraints applied in the model. In a first approach, they only made use of the information regarding whether the leading voice is present or not in each frame. In

a second approach, they took advantage of both time and pitch information on the aligned score.

Zhao *et al.* proposed a score-informed leading voice separation system with a weighting scheme [204]. They extended the system in [202], which is based on the source-filter NMF model in [201], by using a Laplacian or a Gaussian-based mask on the NMF activation matrix to enhance the likelihood of the score-informed pitch candidates.

Jointly estimating accompaniment and lead allowed for some research in correctly estimating the unvoiced parts of the lead, which is the main issue with purely harmonic models, as highlighted in Section III-C. In [201], [205], Durrieu *et al.* extended their model to account for the unvoiced parts by adding white noise components to the voice model.

In the same direction, Janer and Marxer proposed to separate unvoiced fricative consonants using a semi-supervised NMF [206]. They extended the source-filter NMF model in [201] using a low-latency method with timbre classification to estimate the predominant pitch [87]. They approximated the fricative consonants as an additive wideband component, training a model of NMF bases. They also used the transient quality to differentiate between fricatives and drums, after extracting transient time points using the method in [207].

Similarly, Marxer and Janer then proposed to separately model the singing voice breathiness [208]. They estimated the breathiness component by approximating the voice spectrum as a filtered composition of a glottal excitation and a wideband component. They modeled the magnitude of the voice spectrum using the model in [209] and the envelope of the voice excitation using the model in [192]. They estimated the pitch using the method in [87]. This was all integrated into the source-filter NMF model.

The body of research initiated by Durrieu *et al.* in [188] consists of using algebraic models more sophisticated than one simple matrix product, but rather inspired by musicological knowledge. Ozerov *et al.* formalized this idea through a general framework and showed its application for singing voice separation [210]–[212].

Finally, Hennequin and Rigaud augmented their model to account for long-term reverberation, with application to singing voice separation [213]. They extended the model in [214] which allows extraction of the reverberation of a specific source with its dry signal. They combined this model with the source-filter NMF model in [189].

### D. Different Constraints for Different Sources

Algebraic methods that decompose the mixture spectrogram as the sum of the lead and accompaniment spectrograms are based on the minimization of a *cost* or *loss function* which measures the error between the approximation and the observation. While the methods presented above for lead and accompaniment separation did propose more sophisticated models with parameters explicitly pertaining to the lead or the accompaniment, another option that is also popular in the dedicated literature is to modify the cost function of an optimization algorithm for an existing algorithm (e.g., RPCA), so that one part of the resulting components would preferentially account for one source or another.

This approach can be exemplified by the harmonic-percussive source separation method (HPSS), presented in [160], [215], [216]. It consists in filtering a mixture spectrogram so that horizontal lines go in a so-called *harmonic* source, while its vertical lines go into a *percussive* source. Separation is then done with TF masking. Of course, such a method is not adequate for lead and accompaniment separation *per se*, because all the harmonic content of the accompaniment is classified as harmonic. However, it shows that *nonparametric* approaches are also an option, provided the cost function itself is well chosen for each source.

This idea was followed by Yang in [217] who proposed an approach based on RPCA with the incorporation of harmonicity priors and a back-end drum removal procedure to improve the decomposition. He added a regularization term in the algorithm to account for harmonic sounds in the low-rank component and used an NMF-based model trained for drum separation [211] to eliminate percussive sounds in the sparse component.

Jeong and Lee proposed to separate a vocal signal from a music signal [218], extending the HPSS approach in [160], [215]. Assuming that the spectrogram of the signal can be represented as the sum of harmonic, percussive, and vocal components, they derived an objective function which enforces the temporal and spectral continuity of the harmonic and percussive components, respectively, similarly to [160], but also the sparsity of the vocal component. Assuming non-negativity of the components, they then derived iterative update rules to minimize the objective function. Ochiai *et al.* extended this work in [219], notably by imposing harmonic constraints for the lead.

Watanabe *et al.* extended RPCA for singing voice separation [220]. They added a harmonicity constraint in the objective function to account for harmonic structures, such as in vocal signals, and regularization terms to enforce the non-negativity of the solution. They used the generalized forward-backward splitting algorithm [221] to solve the optimization problem. They also applied post-processing to remove the low frequencies in the vocal spectrogram and built a TF mask to remove time frames with low energy.

Going beyond smoothness and harmonicity, Hayashi *et al.* proposed an NMF with a constraint to help separate periodic components, such as a repeating accompaniment [222]. They defined a periodicity constraint which they incorporated in the objective function of the NMF algorithm to enforce the periodicity of the bases.

### E. Cascaded and Iterated Methods

In their effort to propose separation methods for the lead and accompaniment in music, some authors discovered that very different methods often have complementary strengths. This motivated the *combination* of methods. In practice, there are several ways to follow this line of research.

One potential route to achieve better separation is to *cascade* several methods. This is what FitzGerald and Gainza proposed in [216] with multiple median filters [148]. They used a median-filter based HPSS approach at different frequency resolutions to

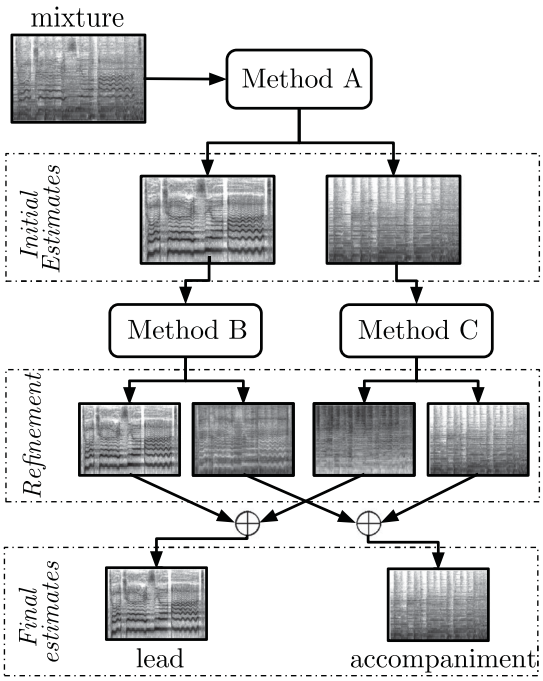



Fig. 8. Cascading source separation methods. The results from method A is improved by applying methods B and C on its output, which are specialized in reducing interferences from undesired sources in each signal.

separate a mixture into harmonic, percussive, and vocal components. They also investigated the use of STFT or CQT as the TF representation and proposed a post-processing step to improve the separation with tensor factorization techniques [223] and non-negative partial co-factorization [180].

The two-stage HPSS system proposed by Tachibana *et al.* in [224] proceeds the same way. It is an extension of the melody extraction approach in [225] and was applied for karaoke in [226]. It consists in using the optimization-based HPSS algorithm from [160], [215], [227], [228] at different frequency resolutions to separate the mixture into harmonic, percussive, and vocal components.

HPSS was not the only separation module considered as the building block of combined lead and accompaniment separation approaches. Deif *et al.* also proposed a multi-stage NMF-based algorithm [229], based on the approach in [230]. They used a local spectral discontinuity measure to refine the non-pitched components obtained from the factorization of the long window spectrogram and a local temporal discontinuity measure to refine the non-percussive components obtained from factorization of the short window spectrogram.

Finally, this cascading concept was considered again by Driedger and Müller in [231], that introduces a processing pipeline for the outputs of different methods [115], [164], [232], [233] to obtain an improved separation quality. Their core idea is depicted in Fig. 8 and combines the output of different methods in a specific order to improve separation.

Another approach for improving the quality of separation when using several separation procedures is not to restrict the number of such iterations from one method to another, but rather to iterate them many times until satisfactory results are obtained.

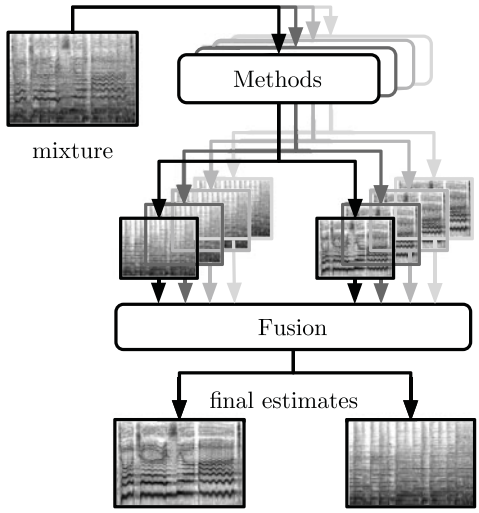



Fig. 9. Fusion of separation methods. The output of many separation methods is fed into a fusion system that combines them to produce a single estimate.

This is what is proposed in Hsu *et al.* in [234], extending the algorithm in [235]. They first estimated the pitch range of the singing voice by using the HPSS method in [160], [225]. They separated the voice given the estimated pitch using a binary mask obtained by training a multilayer perceptron [236] and re-estimated the pitch given the separated voice. Voice separation and pitch estimation are then iterated until convergence.

As another iterative method, Zhu *et al.* proposed a multi-stage NMF [230], using harmonic and percussive separation at different frequency resolutions similar to [225] and [216]. The main originality of their contribution was to iterate the refinements instead of applying it only once.

An issue with such iterated methods lies in how to decide whether convergence is obtained, and it is not clear whether the quality of the separated signals will necessarily improve. For this reason, Bryan and Mysore proposed a user-guided approach based on PLCA, which can be applied for the separation of the vocals [237]–[239]. They allowed a user to make annotations on the spectrogram of a mixture, incorporated the feedback as constraints in a PLCA model [110], [156], and used a posterior regularization technique [240] to refine the estimates, repeating the process until the user is satisfied with the results. This is similar to the way Ozerov *et al.* proposed to take user input into account in [241].

A principled way to aggregate the result of many source separation systems to obtain one single estimate that is consistently better than all of them was presented by Jaureguiberry *et al.* in their *fusion framework*, depicted in Fig. 9. It takes advantage of multiple existing approaches, and demonstrated its application to singing voice separation [242]–[244]. They investigated fusion methods based on non-linear optimization, Bayesian model averaging [245], and deep neural networks (DNN).

As another attempt to design an efficient fusion method, McVicar *et al.* proposed in [246] to combine the outputs of RPCA [115], HPSS [216], Gabor filtered spectrograms [247], REPET [130] and an approach based on deep learning [248]. To do this, they used different classification techniques to build

the aggregated TF mask, such as a logistic regression model or a conditional random field (CRF) trained using the method in [249] with time and/or frequency dependencies.

Manilow *et al.* trained a neural network to predict quality of source separation for three source separation algorithms, each leveraging a different cue—repetition, spatialization, and harmonicity/pitch proximity [250]. The method estimates separation quality of the lead vocals for each algorithm, using only the original audio mixture and separated source output. These estimates were used to guide switching between algorithms along time.

### F. Source-Dependent Representations

In the previous section, we stated that some authors considered iterating separation at different frequency resolutions, i.e., using different TF representations [216], [224], [229]. This can be seen as a combination of different methods. However, this can also be seen from another perspective as based on picking specific *representations*.

Wolf *et al.* proposed an approach using rigid motion segmentation, with application to singing voice separation [251], [252]. They introduced harmonic template models with amplitude and pitch modulations defined by a velocity vector. They applied a wavelet transform [253] on the harmonic template models to build an audio image where the amplitude and pitch dynamics can be separated through the velocity vector. They then derived a velocity equation, similar to the optical flow velocity equation used in images [254], to segment velocity components. Finally, they identified the harmonic templates which model different sources in the mixture and separated them by approximating the velocity field over the corresponding harmonic template models.

Yen *et al.* proposed an approach using spectro-temporal modulation features [255], [256]. They decomposed a mixture using a two-stage auditory model which consists of a cochlear module [257] and cortical module [258]. They then extracted spectro-temporal modulation features from the TF units and clustered the TF units into harmonic, percussive, and vocal components using the EM algorithm and resynthesized the estimated signals.

Chan and Yang proposed an approach using an informed group sparse representation [259]. They introduced a representation built using a learned dictionary based on a chord sequence which exhibits group sparsity [260] and which can incorporate melody annotations. They derived a formulation of the problem in a manner similar to RPCA and solved it using the alternating direction method of multipliers [261]. They also showed a relation between their representation and the low-rank representation in [123], [262].

### G. Shortcomings

The large body of literature we reviewed in the preceding sections is concentrated on choosing adequate models for the lead and accompaniment parts of music signals in order to devise effective signal processing methods to achieve separation. From a higher perspective, their common feature is to guide the separation process in a *model-based way*: first, the scientist has some idea regarding characteristics of the lead signal and/or the accompaniment, and then an algorithm is designed to exploit this knowledge for separation.

Model-based methods for lead and accompaniment separation are faced with a common risk that their core assumptions will be violated for the signal under study. For instance, the lead to be separated may not be harmonic but saturated vocals or the accompaniment may not be repetitive or redundant, but rather always changing. In such cases, model-based methods are prone to large errors and poor performance.

## VI. Data-Driven Approaches

A way to address the potential caveats of model-based separation behaving badly in case of violated assumptions is to avoid making assumptions altogether, but rather to let the model be learned from a large and representative database of examples. This line of research leads to *data-driven* methods, for which researchers are concerned about directly estimating a mapping between the mixture and either the TF mask for separating the sources, or their spectrograms to be used for designing a filter.

As may be foreseen, this strategy based on machine learning comes with several challenges of its own. First, it requires considerable amounts of data. Second, it typically requires a high-capacity learner (many tunable parameters) that can be prone to over-fitting the training data and therefore not working well on the audio it faces when deployed.

### A. Datasets

Building a good data-driven method for source separation relies heavily on a training dataset to learn the separation model. In our case, this not only means obtaining a set of musical songs, but also their constitutive accompaniment and lead sources, summing up to the mixtures. For professionally-produced or recorded music, the separated sources are often either unavailable or private. Indeed, they are considered amongst the most precious assets of right holders, and it is very difficult to find isolated vocals and accompaniment of professional bands that are freely available for the research community to work on without copyright infringements.

Another difficulty arises when considering that the different sources in a musical content do share some common orchestration and are not superimposed in a random way, prohibiting simply summing isolated random notes from instrumental databases to produce mixtures. This contrasts with the speech community which routinely generates mixtures by summing noise data [263] and clean speech [264].

Furthermore, the temporal structures in music signals typically spread over long periods of time and can be exploited to achieve better separation. Additionally, short excerpts do not often comprise parts where the lead signal is absent, although a method should learn to deal with that situation. This all suggests that including full songs in the training data is preferable over short excerpts.

Finally, professional recordings typically undergo sophisticated sound processing where panning, reverberation, and other sound effects are applied to each source separately, and also to

TABLE I
SUMMARY OF DATASETS AVAILABLE FOR LEAD AND ACCOMPANIMENT SEPARATION

| Dataset | Year | Reference(s) | URL | Tracks | Track duration (s) | Full/stereo? |
|---|---|---|---|---|---|---|
| MASS | 2008 | [266] | http://www.mtg.upf.edu/download/datasets/mass | 9 | $16 \pm 7$ | no / yes |
| MIR-1K | 2010 | [74] | https://sites.google.com/site/unvoicedsoundseparation/mir-1k | 1,000 | $8 \pm 8$ | no / no |
| QUASI | 2011 | [270], [273] | http://www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/ | 5 | $206 \pm 21$ | yes / yes |
| ccMixter | 2014 | [141] | http://www.loria.fr/~aliutkus/kam/ | 50 | $231 \pm 77$ | yes / yes |
| MedleyDB | 2014 | [274] | http://medleydb.weebly.com/ | 63 | $206 \pm 121$ | yes / yes |
| iKala | 2015 | [162] | http://mac.citi.sinica.edu.tw/ikala/ | 206 | 30 | no / no |
| DSD100 | 2015 | [271] | sisec17.audiolabs-erlangen.de | 100 | $251 \pm 60$ | yes / yes |
| MUSDB18 | 2017 | [275] | https://sigsep.github.io/musdb | 150 | $236 \pm 95$ | yes / yes |

Tracks without vocals were omitted in the statistics.

the mixture. To date, simulated data sets have poorly mimicked these effects [265]. Many separation methods make assumptions about the mixing model of the sources, e.g., assuming it is linear (i.e., does not comprise effects such as dynamic range compression). It is quite common that methods giving extremely good performance for linear mixtures completely break down when processing published musical recordings. Training and test data should thus feature realistic audio engineering to be useful for actual applications.

In this context, the development of datasets for lead and accompaniment separation was a long process. In early times, it was common for researchers to test their methods on some private data. To the best of our knowledge, the first attempt at releasing a public dataset for evaluating vocals and accompaniment separation was the Music Audio Signal Separation (MASS) dataset [266]. It strongly boosted research in the area, even if it only featured 2.5 minutes of data. The breakthrough was made possible by some artists which made their mixed-down audio, as well as its constitutive stems (unmixed tracks), available under open licenses such as Creative Commons, or authorized scientists to use their material for research.

The MASS dataset then formed the core content of the early Signal Separation Evaluation Campaigns (SiSEC) [267], which evaluate the quality of various music separation methods [268]–[272]. SiSEC always had a strong focus on vocals and accompaniment separation. For a long time, vocals separation methods were very demanding computationally and it was already considered extremely challenging to separate excerpts of only a few seconds.

In the following years, new datasets were proposed that improved over the MASS dataset in many directions. We briefly describe the most important ones, summarized in Table I.

- The QUASI dataset was proposed to study the impact of different mixing scenarios on the separation quality. It consists of the same tracks as in the MASS dataset, but kept full length and mixed by professional sound engineers.
- The MIR-1K and iKala datasets were the first attempts to scale vocals separation up. They feature a higher number of samples than the previously available datasets. However, they consist of mono signals of very short and amateur karaoke recordings.
- The ccMixter dataset was proposed as the first dataset to feature many full-length stereo tracks. Each one comes with a vocals and an accompaniment source. Although it is stereo, it often suffers from simplistic mixing of sources, making it unrealistic in some aspects.
- MedleyDB has been developed as a dataset to serve many purposes in music information retrieval. It consists of more than 100 full-length recordings, with all their constitutive sources. It is the first dataset to provide such a large amount of data to be used for audio separation research (more than 7 hours). Among all the material present in that dataset, 63 tracks feature singing voice.
- DSD100 was presented for SiSEC 2016. It features 100 full-length tracks originating from the 'Mixing Secret' Free Multitrack Download Library[1] of the Cambridge Music Technology, which is freely usable for research and educational purposes.

Finally, we present here the MUSDB18 dataset, putting together tracks from MedleyDB, DSD100, and other new musical material. It features 150 full-length tracks, and has been constructed by the authors of this paper so as to address all the limitations we identified above:

- It only features full-length tracks, so that the handling of long-term musical structures, and of silent regions in the lead/vocal signal, can be evaluated.
- It only features stereo signals which were mixed using professional digital audio workstations. This results in quality stereo mixes which are representative of real application scenarios.
- As with DSD100, a design choice of MUSDB18 was to split the signals into 4 predefined categories: bass, drums, vocals, and other. This contrasts with the enhanced granularity of MedleyDB that offers more types of sources, but it strongly promotes automation of the algorithms.
- Many musical genres are represented in MUSDB18, for example, jazz, electro, metal, etc.
- It is split into a development (100 tracks, 6.5 h) and a test dataset (50 tracks, 3.5 h), for the design of data-driven separation methods.

All details about this freely available dataset and its accompanying software tools may be found in its dedicated website.[2]

In any case, it can be seen that datasets of sufficient duration to build data-driven separation methods were only created recently.

### B. Algebraic Approaches

A natural way to exploit a training database was to learn some parts of the model to guide the estimation process into better solutions. Work on this topic may be traced back to the suggestion

[1] http://www.cambridge-mt.com/ms-mtk.htm
[2] https://sigsep.github.io/musdb

of Ozerov *et al.* in [276] to learn spectral template models based on a database of isolated sources, and then to adapt this dictionary of templates on the mixture using the method in [277].

The exploitation of training data was formalized by Smaragdis *et al.* in [110] in the context of source separation within the supervised and semi-supervised PLCA framework. The core idea of this probabilistic formulation, equivalent to NMF, is to learn some spectral bases from the training set which are then kept fixed at separation time.

In the same line, Ozerov *et al.* proposed an approach using Bayesian models [191]. They first segmented a song into vocal and non-vocal parts using GMMs with MFCCs. Then, they adapted a general music model on the non-vocal parts of a particular song by using the maximum a posteriori (MAP) adaptation approach in [278].

Ozerov *et al.* later proposed a framework for source separation which generalizes several approaches given prior information about the problem and showed its application for singing voice separation [210]–[212]. They chose the local Gaussian model [279] as the core of the framework and allowed the prior knowledge about each source and its mixing characteristics using user-specified constraints. Estimation was performed through a generalized EM algorithm [32].

Rafii *et al.* proposed in [280] to address the main drawback of the repetition-based methods described in Section IV-C, which is the weakness of the model for vocals. For this purpose, they combined the REPET-SIM model [135] for the accompaniment with a NMF-based model for singing voice learned from a voice dataset.

As yet another example of using training data for NMF, Boulanger-Lewandowski *et al.* proposed in [281] to exploit long-term temporal dependencies in NMF, embodied using recurrent neural networks (RNN) [236]. They incorporated RNN regularization into the NMF framework to temporally constrain the activity matrix during the decomposition, which can be seen as a generalization of the non-negative HMM in [282]. Furthermore, they used supervised and semi-supervised NMF algorithms on isolated sources to train the models, as in [110].

### C. Deep Neural Networks

Taking advantage of the recent availability of sufficiently large databases of isolated vocals along with their accompaniment, several researchers investigated the use of machine learning methods to directly estimate a mapping between the mixture and the sources. Although end-to-end systems inputting and outputting the waveforms have already been proposed in the speech community [283], they are not yet available for music source separation. This may be due to the relative small size of music separation databases, at most 10 h today. Instead, most systems feature pre and post-processing steps that consist in computing classical TF representations and building TF masks, respectively. Although such end-to-end systems will inevitably be proposed in the near future, the common structure of deep learning methods for lead and accompaniment separation usually corresponds for now to the one depicted in Fig. 10. From a general perspective, we may say that most current methods

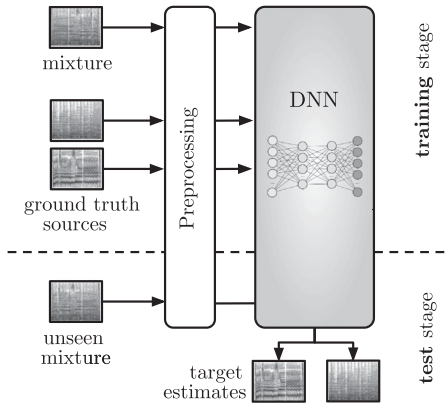



Fig. 10. General architecture for methods exploiting deep learning. The network inputs the mixture and outputs either the sources spectrograms or a TF mask. Methods usually differ in their choice for a network architecture and the way it is learned using the training data.

mainly differ in the structure picked for the network, as well as in the way it is learned.

Providing a thorough introduction to deep neural networks is out of the scope of this paper. For our purpose, it suffices to mention that they consist of a cascade of several possibly non-linear transformations of the input, which are learned during a training stage. They were shown to effectively learn representations and mappings, provided enough data is available for estimating their parameters [284]–[286]. Different architectures for neural networks may be combined/cascaded together, and many architectures were proposed in the past, such as feedforward fully-connected neural networks (FNN), convolutional neural networks (CNN), or RNN and variants such as the long short-term memory (LSTM) and the gated-recurrent units (GRU). Training of such functions is achieved by stochastic gradient descent [287] and associated algorithms, such as backpropagation [288] or backpropagation through time [236] for the case of RNNs.

To the best of our knowledge, Huang *et al.* were the first to propose deep neural networks, RNNs here [289], [290], for singing voice separation in [248], [291]. They adapted their framework from [292] to model all sources simultaneously through masking. Input and target functions were the mixture magnitude and a joint representation of the individual sources. The objective was to estimate jointly either singing voice and accompaniment music, or speech and background noise from the corresponding mixtures.

Modeling the temporal structures of both the lead and the accompaniment is a considerable challenge, even when using DNN methods. As an alternative to the RNN approach proposed by Huang *et al.* in [248], Uhlich *et al.* proposed the usage of FNNs [293] whose input consists of *supervectors* of a few consecutive frames from the mixture spectrogram. Later in [294], the same authors considered the use of bi-directional LSTMs for the same task.

In an effort to make the resulting system less computationally demanding at separation time but still incorporating dynamic modeling of audio, Simpson *et al.* proposed in [295] to predict binary TF masks using deep CNNs, which typically utilize

fewer parameters than the FNNs. Similarly, Schlueter proposed a method trained to detect singing voice using CNNs [296]. In that case, the trained network was used to compute *saliency maps* from which TF masks can be computed for singing voice separation. Chandna *et al.* also considered CNNs for lead separation in [297], with a particular focus on low-latency.

The classical FNN, LSTM and CNN structures above served as baseline structures over which some others tried to improve. As a first example, Mimilakis *et al.* proposed to use a hybrid structure of FNNs with skip connections to separate the lead instrument for purposes of remixing jazz recordings [298]. Such skip connections allow to propagate the input spectrogram to intermediate representations within the network, and mask it similarly to the operation of TF masks. As advocated, this enforces the networks to approximate a TF masking process. Extensions to temporal data for singing voice separation were presented in [299], [300]. Similarly, Jansson *et al.* proposed to propagate the spectral information computed by convolutional layers to intermediate representations [301]. This propagation aggregates intermediate outputs to proceeding layer(s). The output of the last layer is responsible for masking the input mixture spectrogram. In the same vein, Takahashi *et al.* proposed to use skip connections via element-wise addition through representations computed by CNNs [302].

Apart from the structure of the network, the way it is trained, comprising how the targets are computed, has a tremendous impact on performance. As we saw, most methods operate on defining TF masks or estimating magnitude spectrograms. However, other methods were proposed based on deep clustering [303], [304], where TF mask estimation is seen as a clustering problem. Luo *et al.* investigated both approaches in [305] by proposing deep bidirectional LSTM networks capable of outputting both TF masks or features to use as in deep clustering. Kim and Smaragdis proposed in [306] another way to learn the model, in a denoising auto-encoding fashion [307], again utilizing short segments of the mixture spectrogram as an input to the network, as in [293].

As the best network structure may vary from one track to another, some authors considered a fusion of methods, in a manner similar to the method [242] presented above. Grais *et al.* [308], [309] proposed to aggregate the results from an ensemble of feedforward DNNs to predict TF masks for separation. An improvement was presented in [310], [311] where the inputs to the fusion network were separated signals, instead of TF masks, aiming at enhancing the reconstruction of the separated sources.

As can be seen the use of deep learning methods for the design of lead and accompaniment separation has already stimulated a lot of research, although it is still in its infancy. Interestingly, we also note that using audio and music specific knowledge appears to be fundamental in designing effective systems. As an example of this, the contribution from Nie *et al.* in [312] was to include the construction of the TF mask as an extra non-linearity included in a recurrent network. This is an exemplar of where signal processing elements, such as filtering through masking, are incorporated as a building block of the machine learning method.

The network structure is not the only thing that can benefit from audio knowledge for better separation. The design of appropriate features is another. While we saw that supervectors of spectrogram patches offered the ability to effectively model time-context information in FNNs [293], Sebastian and Murthy [313] proposed the use of the modified group delay feature representation [314] in their deep RNN architecture. They applied their approach for both singing voice and vocal-violin separation.

Finally, as with other methods, DNN-based separation techniques can also be combined with others to yield improved performance. As an example, Fan *et al.* proposed to use DNNs to separate the singing voice and to also exploit vocal pitch estimation [315]. They first extracted the singing voice using feedforward DNNs with sigmoid activation functions. They then estimated the vocal pitch from the extracted singing voice using dynamic programming.

### D. Shortcomings

Data-driven methods are nowadays the topic of important research efforts, particularly those based on DNNs. This is notably due to their impressive performance in terms of separation quality, as can, for instance, be noticed below in Section VIII. However, they also come with some limitations.

First, we highlighted that lead and accompaniment separation in music has the very specific problem of scarce data. Since it is very hard to gather large amounts of training data for that application, it is hard to fully exploit learning methods that require large training sets. This raises very specific challenges in terms of machine learning.

Second, the lack of interpretability of model parameters is often mentioned as a significant shortcoming when it comes to applications. Indeed, music engineering systems are characterized by a strong importance of human-computer interactions, because they are used in an artistic context that may require specific needs or results. As of today, it is unclear how to provide user interaction for controlling the millions of parameters of DNN-based systems.

## VII. Including Multichannel Information

In describing the above methods, we have not discussed the fact that music signals are typically stereophonic. On the contrary, the bulk of methods we discussed focused on designing good spectrogram models for the purpose of filtering mixtures that may be *monophonic*. Such a strategy is called *single-channel* source separation and is usually presented as more challenging than multichannel source separation. Indeed, only TF structure may then be used to discriminate the accompaniment from the lead. In stereo recordings, one further so-called *spatial* dimension is introduced, which is sometimes referred to as *pan*, that corresponds to the perceived *position* of a source in the stereo field. Devising methods to exploit this spatial diversity for source separation has also been the topic of an important body of research that we review now.

### A. Extracting the Lead based on Panning

In the case of popular music signals, a fact of paramount practical importance is that the lead signal—such as vocals—is very often mixed *in the center*, which means that its energy

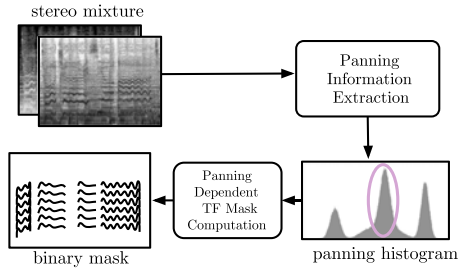



Fig. 11. Separation of the lead based on panning information. A stereo cue called panning allows to design a TF mask.

is approximately the same in left and right channels. On the contrary, other instruments are often mixed at positions to the left or right of the stereo field.

The general structure of methods extracting the lead based on stereo cues is displayed on Fig. 11, introduced by Avendano, who proposed to separate sources in stereo mixtures by using a panning index [316]. He derived a two-dimensional map by comparing left and right channels in the TF domain to identify the different sources based on their panning position [317]. The same methodology was considered by Barry *et al.* in [318] in his Azimuth Discrimination and Resynthesis (ADRess) approach, with panning indexes computed with differences instead of ratios.

Vinyes *et al.* also proposed to unmix commercially produced music recordings thanks to stereo cues [319]. They designed an interface similar to [318] where a user can set some parameters to generate different TF filters in real time. They showed applications for extracting various instruments, including vocals.

Cobos and López proposed to separate sources in stereo mixtures by using TF masking and multilevel thresholding [320]. They based their approach on the Degenerate Unmixing Estimation Technique (DUET) [321]. They first derived histograms by measuring the amplitude relationship between TF points in left and right channels. Then, they obtained several thresholds using the multilevel extension of Otsu's method [322]. Finally, TF points were assigned to their related sources to produce TF masks.

Sofianos *et al.* proposed to separate the singing voice from a stereo mixture using ICA [323]–[325]. They assumed that most commercial songs have the vocals panned to the center and that they dominate the other sources in amplitude. In [323], they proposed to combine a modified version of ADRess with ICA to filter out the other instruments. In [324], they proposed a modified version without ADRess.

Kim *et al.* proposed to separate centered singing voice in stereo music by exploiting binaural cues, such as inter-channel level and inter-channel phase difference [326]. To this end, they build the pan-based TF mask through an EM algorithm, exploiting a GMM model on these cues.

### *B. Augmenting Models With Stereo*

As with using only a harmonic model for the lead signal, using stereo cues in isolation is not always sufficient for good separation, as there can often be multiple sources at the same spatial location. Combining stereo cues with other methods improves performance in these cases.

Cobos and López proposed to extract singing voice by combining panning information and pitch tracking [327]. They first obtained an estimate for the lead thanks to a pan-based method such as [316], and refined the singing voice by using a TF binary mask based on comb-filtering method as in Section III-B. The same combination was proposed by Marxer *et al.* in [87] in a low-latency context, with different methods used for the binaural cues and pitch tracking blocks.

FitzGerald proposed to combine approaches based on repetition and panning to extract stereo vocals [328]. He first used his nearest neighbors median filtering algorithm [139] to separate vocals and accompaniment from a stereo mixture. He then used the ADRess algorithm [318] and a high-pass filter to refine the vocals and improve the accompaniment. In a somewhat different manner, FitzGerald and Jaiswal also proposed to combine approaches based on repetition and panning to improve stereo accompaniment recovery [329]. They presented an audio inpainting scheme [330] based on the nearest neighbors and median filtering algorithm [139] to recover TF regions of the accompaniment assigned to the vocals after using a source separation algorithm based on panning information.

In a more theoretically grounded manner, several methods based on a probabilistic model were generalized to the multichannel case. For instance, Durrieu *et al.* extended their source-filter model in [201], [205] to handle stereo signals, by incorporating the panning coefficients as model parameters to be estimated.

Ozerov and Févotte proposed a multichannel NMF framework with application to source separation, including vocals and music [331], [332]. They adopted a statistical model where each source is represented as a sum of Gaussian components [193], and where maximum likelihood estimation of the parameters is equivalent to NMF with the Itakura-Saito divergence [94]. They proposed two methods for estimating the parameters of their model, one that maximized the likelihood of the multichannel data using EM, and one that maximized the sum of the likelihoods of all channels using a multiplicative update algorithm inspired by NMF [90].

Ozerov *et al.* then proposed a multichannel non-negative tensor factorization (NTF) model with application to user-guided source separation [333]. They modeled the sources jointly by a 3-valence tensor (time/frequency/source) as in [334] which extends the multichannel NMF model in [332]. They used a generalized EM algorithm based on multiplicative updates [335] to minimize the objective function. They incorporated information about the temporal segmentation of the tracks and the number of components per track. Ozerov *et al.* later proposed weighted variants of NMF and NTF with application to user-guided source separation, including separation of vocals and music [241], [336].

Sawada *et al.* also proposed multichannel extensions of NMF, tested for separating stereo mixtures of multiple sources, including vocals and accompaniment [337]–[339]. They first defined multichannel extensions of the cost function, namely, Euclidean distance and Itakura-Saito divergence, and derived multiplicative update rules accordingly. They then proposed

two techniques for clustering the bases, one built into the NMF model and one performing sequential pair-wise merges.

Finally, multichannel information was also used with DNN models. Nugraha *et al.* addressed the problem of multichannel source separation for speech enhancement [340], [341] and music separation [342], [343]. In this framework, DNNs are still used for the spectrograms, while more classical EM algorithms [344], [345] are used for estimating the spatial parameters.

### C. Shortcomings

When compared to simply processing the different channels independently, incorporating spatial information in the separation method often comes at the cost of additional computational complexity. The resulting methods are indeed usually more demanding in terms of computing power, because they involve the design of beamforming filters and inversion of covariance matrices. While this is not really an issue for stereophonic music, this may become prohibiting in configurations with higher numbers of channels.

## VIII. Evaluation

### A. Background

The problem of evaluating the quality of audio signals is a research topic of its own, which is deeply connected to psychoacoustics [346] and has many applications in engineering because it provides an objective function to optimize when designing processing methods. While mean squared error (MSE) is often used for mathematical convenience whenever an error is to be computed, it is a very established fact that MSE is not representative of audio perception [347], [348]. For example, inaudible phase shifts would dramatically increase the MSE. Moreover, it should be acknowledged that the concept of quality is rather application-dependent.

In the case of signal separation or enhancement, processing is often only a part of a whole architecture and a relevant methodology for evaluation is to study the positive or negative impact of this module on the overall performance of the system, rather than to consider it independently from the rest. For example, when embedded in an automatic speech recognition (ASR) system, performance of speech denoising can be assessed by checking whether it decreases word error rate [349].

When it comes to music processing, and more particularly to lead and accompaniment separation, the evaluation of separation quality has traditionally been inspired by work in the audio coding community [347], [350] in the sense that it aims at comparing ground truth vocals and accompaniment with their estimates, just like audio coding compares the original with the compressed signal.

### B. Metrics

As noted previously, MSE-based error measures are not perceptually relevant. For this reason, a natural approach is to have humans do the comparison. The gold-standard for human perceptual studies is the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) methodology, that is commonly used for evaluating audio coding [350].

However, it quickly became clear that the specific evaluation of separation quality cannot easily be reduced to a single number, even when achieved through actual perceptual campaigns, but that quality rather depends on the application considered. For instance, karaoke or vocal extraction come with opposing trade-offs between isolation and distortion. For this reason, it has been standard practice to provide different and complementary metrics for evaluating separation that measure the amount of distortion, artifacts, and interference in the results.

While human-based perceptual evaluation is definitely the best way to assess separation quality [351], [352], having computable objective metrics is desirable for several reasons. First, it allows researchers to evaluate performance without setting up costly and lengthy perceptual evaluation campaigns. Second, it permits large-scale training for the fine-tuning of parameters. In this respect, the Blind Source Separation Evaluation (BSS Eval) toolbox [353], [354] provides quality metrics in decibel to account for distortion (SDR), artifacts (SAR), and interferences (SIR). Since it was made available quite early and provides somewhat reasonable correlation with human perception in certain cases [355], [356] it is still widely used to this day.

Even if BSS Eval was considered sufficient for evaluation purposes for a long time, it is based on squared error criteria. Following early work in the area [357], the Perceptual Evaluation of Audio Source Separation (PEASS) toolkit [358]–[360] was introduced as a way to predict perceptual ratings. While the methodology is very relevant, PEASS however was not widely accepted in practice. We believe this is for two reasons. First, the proposed implementation is quite computationally demanding. Second, the perceptual scores it was designed with are more related to speech separation than to music.

Improving perceptual evaluation often requires a large amount of experiments, which is both costly and requires many expert listeners. One way to increase the number of participants is to conduct web-based experiments. In [361], the authors report they were able to aggregate 530 participants in only 8.2 hours and obtained perceptual evaluation scores comparable to those estimated in the controlled lab environment.

Finally, we highlight here that the development of new perceptually relevant objective metrics for singing voice separation evaluation remains an open issue [362]. It is also a highly crucial one for future research in the domain.

### C. Performance (SiSEC 2016)

In this section, we will discuss the performance of 23 source separation methods evaluated on the DSD100, as part of the task for separating professionally-produced music recordings at SiSEC 2016. The methods are listed in Table II, along with the acronyms we use for them, their main references, a very brief summary, and a link to the section where they are described in the text. To date, this stands as the largest evaluation campaign ever achieved on lead and accompaniment separation. The results we discuss here are a more detailed report for SiSEC 2016 [272], presented in line with the taxonomy proposed in this paper.

TABLE II
METHODS EVALUATED DURING SISEC 2016

| Acronym | Ref. | Summary | Section |
|---|---|---|---|
| HUA | [115] | RPCA standard version | IV-B |
| RAF1 | [130] | REPET standard version | IV-C |
| RAF2 | [134] | REPET with time-varying period | |
| RAF3 | [135] | REPET with similarity matrix | |
| KAM1-2 | [142] | KAM with different configurations | |
| CHA | [162] | RPCA with vocal activation information | V-A |
| JEO1-2 | [163] | $l_1$-RPCA with vocal activation information | |
| DUR | [201] | Source-filter NMF | V-C |
| OZE | [212] | Structured NMF with learned dictionaries | VI-B |
| KON | [291] | RNN | VI-C |
| GRA2-3 | [308] | DNN ensemble | |
| STO1-2 | [363] | FNN on *common fate* TF representation | |
| UHL1 | [293] | FNN with context | |
| NUG1-4 | [343] | FNN with multichannel information | VII |
| UHL2-3 | [294] | LSTM with multichannel information | |
| IBM | | ideal binary mask | |

The objective scores for these methods were obtained using BSS Eval and are given in Fig. 12. For more details about the results and for listening to the estimates, we refer the reader to the dedicated interactive website.[3]

As we first notice in Fig. 12, the HUA method, corresponding to the standard RPCA as discussed in Section IV-B, showed rather disappointing performance in this evaluation. After inspection of the results, it appears that processing full-length tracks is the issue there: at such scales, vocals also exhibit redundancy, which is captured by the low-rank model associated with the accompaniment. On the other hand, the RAF1-3 and KAM1-3 methods that exploit redundancy through repetitions, as presented in Section IV-C, behave much better for full-length tracks: even if somewhat redundant, vocals are rarely as repetitive as the accompaniment. When those methods are evaluated on datasets with very short excerpts (e.g., MIR-1K), such severe practical drawbacks are not apparent.

Likewise, the DUR method that jointly models the vocals as harmonic and the accompaniment as redundant, as discussed in Section V-C, does show rather disappointing performance, considering that it was long the state-of-the-art in earlier SiSECs [270]. After inspection, we may propose two reasons for this performance drop. First, using full-length excerpts also clearly revealed a shortcoming of the approach: it poorly handles silences in the lead, which were rare in the short-length excerpts tested so far. Second, using a much larger evaluation set revealed that vocals are not necessarily well modeled by a harmonic source-filter model; breathy or saturated voices appear to greatly challenge such a model.

While processing full-length tracks comes as a challenge, it can also be an opportunity. It is indeed worth noticing that whenever RPCA is helped through vocal activity detection, its performance is significantly boosted, as highlighted by the relatively good results obtained by CHAN and JEO.

As discussed in Section VI, the availability of learning data made it possible to build data-driven approaches, like the NMF-based OZE method which is available through the Flexible Audio Source Separation Toolbox (FASST) [211], [212]. Although it was long state-of-the-art, it has been strongly outperformed recently by other data-driven approaches, namely DNNs. One first reason clearly appears as the superior expressive power of DNNs over NMF, but one second reason could very simply be that OZE should be trained anew with the same large amount of data.

As mentioned above, a striking fact we see in Fig. 12 is that the overall performance of data-driven DNN methods is the highest. This shows that exploiting learning data does help separation greatly compared to only relying on *a priori* assumptions such as the harmonicity or redundancy. Additionally, dynamic models such as CNN or LSTM appear more adapted to music than FNN. These good performances in audio source separation go in line with the recent success of DNNs in fields as varied as computer vision, speech recognition, and natural language processing [285].

However, the picture may be seen to be more subtle than simply black-box DNN systems beating all other approaches. For instance, exploiting multichannel probabilistic models, as discussed in Section VII, leads to the NUG and UHL2-3 methods, that significantly outperform the DNN methods ignoring stereo information. In the same vein, we expect other specific assumptions and musicological ideas to be exploited for further improving the quality of the separation.

One particular feature of this evaluation is that it also shows obvious weaknesses in the objective metrics. For instance, the GRA method behaves significantly worse than any other methods. However, when listening to the separated signals, this does not seem deserved. All in all, designing new and convenient metrics that better match perception and that are specifically built for music on large datasets clearly appears as a desirable milestone.

In any case, the performance achieved by a totally informed filtering method such as IBM is significantly higher than that of any submitted method in this evaluation. This means that lead and accompaniment separation has room for much improvement, and that the topic is bound to witness many breakthroughs still. This is even more true considering that IBM is not the best upper bound for separation performance: other filtering methods such as *ideal ratio mask* [20] or multichannel Wiener filter [344] may be considered as references.

Regardless of the above, we would also like to highlight that good algorithms and models can suffer from slight errors in their low-level audio processing routines. Such routines may include the STFT representation, the overlap-add procedure, energy normalization, and so on. Considerable improvements may also be obtained by using simple tricks and, depending on the method, large impacts can occur in the results by only changing low-level parameters. These include the overlap ratio for the STFT, specific ways to regularize matrix inverses in multichannel models, etc. Further tricks such as the exponentiation of the TF mask by some positive value can often boost performance significantly more than using more sophisticated models. However, such tricks are often lost when publishing research focused on the higher-level algorithms. We believe this is an important reason why sharing source code is highly desirable in this particular application. Some online repositories

[3]http://www.sisec17.audiolabs-erlangen.de

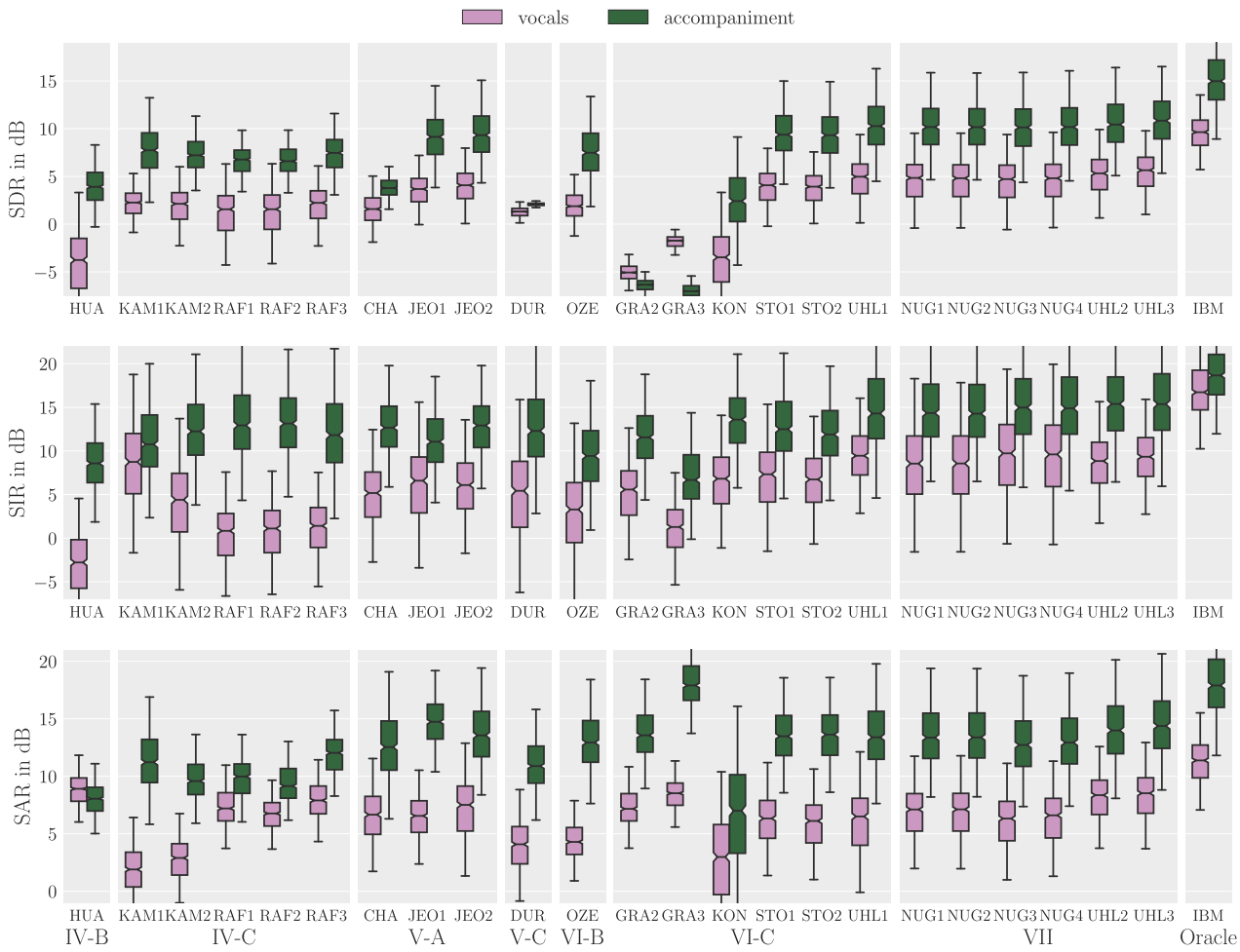



Fig. 12. BSS Eval scores for the vocals and accompaniment estimates for SiSEC 2016 on the DSD100 dataset. Results are shown for the *test* set only. Scores are grouped as in Table II according to the section they are described in the text, indicated below each group.

containing implementations of lead and accompaniment separation methods should be mentioned, such as **nussl**[4] and **untwist** [364]. In the companion webpage of this paper,[5] we list many different online resources such as datasets, implementations, and tools that we hope will be useful to the practitioner and provide some useful pointers to the interested reader.

### D. Discussion

Finally, we summarize the core advantages and disadvantages for each one of the five groups of methods we identified.

Methods based on the harmonicity assumption for the lead are focused on sinusoidal modeling. They enjoy a very strong interpretability and allow for the direct incorporation of any prior knowledge concerning pitch. Their fundamental weakness lies in the fact that many singing voice signals are not harmonic, e.g., when breathy or distorted.

Modeling the accompaniment as redundant allows to exploit long-term dependencies in music signals and may benefit from high-level information like tempo or score. Their most important drawback is to fall short in terms of voice models: the lead signal itself is often redundant to some extent and thus partly incorporated in the estimated accompaniment.

Systems jointly modeling the lead as harmonic and the accompaniment as redundant benefit from both assumptions. They were long state-of-the-art and enjoy a good interpretability, which makes them good candidates for interactive separation methods. However, their core shortcoming is to be highly sensitive to violations of their assumptions, which proves to often be the case in practice. Such situations usually require fine-tuning and hence prevents their use as black-box systems for a broad audience.

Data-driven methods involve machine learning to directly learn a mapping between the mixture and the constitutive sources. Such a strategy recently introduced a breakthrough compared to everything that was done before. Their most important disadvantages are the lack of interpretability, which makes it challenging to design good user interactions, as well as their strong dependency on the size of the training data.

Finally, multichannel methods leverage stereophonic information to strongly improve performance. Interestingly, this can usually be combined with better spectrogram models such as

[4]https://github.com/interactiveaudiolab/nussl
[5]https://sigsep.github.io

DNNs to further improve quality. The price to pay for this boost in performance is an additional computational cost, that may be prohibitive for recordings of more than two channels.

## IX. Conclusion

In this paper, we thoroughly discussed the problem of separating lead and accompaniment signals in music recordings. We gave a comprehensive overview of the research undertaken in the last 50 years on this topic, classifying the different approaches according to their main features and assumptions. In doing so, we showed how one very large body of research can be described as being model-based. In this context, it was evident from the literature that the two most important assumptions behind these models are that the lead instrument is harmonic, while the accompaniment is redundant. As we demonstrated, a very large number of methods on model-based lead-accompaniment separation can be seen as using one or both of these assumptions.

However, music encompasses a variety of signals of an extraordinary diversity, and no rigid assumption holds well for all signals. For this reason, while there are often some music pieces where each method performs well, there will also be some where it fails. As a result, data-driven methods were proposed as an attempt to introduce more flexibility at the cost of requiring representative training data. In the context of this paper, we proposed the largest freely available dataset for music separation, comprising close to 10 hours of data, which is 240 times greater than the first public dataset released 10 years ago.

At present, we see a huge focus on research utilizing recent machine learning breakthroughs for the design of singing voice separation methods. This came with an associated boost in performance, as measured by objective metrics. However, we have also discussed the strengths and shortcomings of existing evaluations and metrics. In this respect, it is important to note that the songs used for evaluation are but a minuscule fraction of all recorded music, and that separating music signals remains the processing of an artistic means of expression. As such it is impossible to escape the need for human perceptual evaluations, or at least adequate models for it.

After reviewing the large existing body of literature, we may conclude here by saying that lead and accompaniment separation in music is a problem at the crossroads of many different paradigms and methods. Researchers from very different backgrounds such as physics, signal or computer engineering have tackled it, and it exists both as an area for strong theoretical research and as a real-world challenging engineering problem. Its strong connections with the arts and digital humanities have proved attractive to many researchers.

Finally, as we showed, there is still much room for improvement in lead and accompaniment separation, and we believe that new and exciting research will bring new breakthroughs in this field. While DNN methods represent the latest big step forward and significantly outperform previous research, we believe that future improvements can come from any direction, including those discussed in this paper. Still, we expect future improvements to initially come from improved machine learning methodologies that can cope with reduced training sets, as well as improved modeling of the specific properties of musical signals, and the development of better signal representations.

## References


[1] R. Kalakota and M. Robinson, *e-Business 2.0: Roadmap for Success*. Boston, MA, USA: Addison-Wesley, 2000.

[2] C. K. Lam and B. C. Tan, "The Internet is changing the music industry," *Commun. ACM*, vol. 44, no. 8, pp. 62–68, 2001.

[3] P. Common and C. Jutten, *Handbook of Blind Source Separation*. New York, NY, USA: Academic, 2010.

[4] G. R. Naik and W. Wang, *Blind Source Separation*. Berlin, Germany: Springer, 2014.

[5] A. Hyvärinen, "Fast and robust fixed-point algorithm for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999.

[6] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4/5, pp. 411–430, Jun. 2000.

[7] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. Dordrecht, The Netherlands: Springer, 2007.

[8] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. New York, NY, USA: Wiley, 2018.

[9] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 1990.

[10] A. Liutkus, J.-L. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *Proc. 14th Int. Workshop Image Anal. Multimedia Interact. Serv.*, Paris, France, Jul. 2013.

[11] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 107–115, May 2014.

[12] U. Zölzer, *DAFX - Digital Audio Effects*. New York, NY, USA: Wiley, 2011.

[13] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. New York, NY, USA: Springer, 2015.

[14] E. T. Jaynes, *Probability Theory: The logic of Science*. Cambridge U.K.: Cambridge Univ. Press, 2003.

[15] O. Cappé, E. Moulines, and T. Ryden, *Inference in Hidden Markov Models (Springer Series in Statistics)*. Secaucus, NJ, USA: Springer, 2005.

[16] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.

[17] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, USA, May 2002, pp. I-529–I-532.

[18] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[19] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time.* Cambridge, MA, USA: MIT Press, 1975.

[20] A. Liutkus and R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, Apr. 2015, pp. 266–270.

[21] G. Fant, *Acoustic Theory of Speech Production*. Berlin, Germany: Walter de Gruyter, 1970.

[22] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency alanysis of time series for echoes: Cepstrum pseudo-autocovariance, cross-cepstrum, and saphe cracking," in *Proc. Symp. Time Series Anal.*, 1963, pp. 209–243.

[23] A. M. Noll, "Short-time spectrum and "cepstrum" techniques for vocal-pitch detection," *J. Acoust. Soc. Amer.*, vol. 36, no. 2, pp. 296–302, 1964.

[24] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, no. 2, pp. 293–309, 1967.

[25] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[26] A. V. Oppenheim, "Speech Anal.-synthesis system based on homomorphic filtering," *J. Acoust. Soc. Amer.*, vol. 45, no. 2, pp. 458–465, 1969.

[27] R. Durrett, *Probability: Theory and Examples*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[28] G. Schwarz, "Estimating the dimension of a model," *Annals Stat.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[29] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[30] A. J. Viterbi, "A personal history of the Viterbi algorithm," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 120–142, 2006.
[31] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1996.
[32] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.
[33] J. Salamon, E. Gómez, D. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications and challenges," *IEEE Signal Process. Mag.*, vol. 31, no. 2, pp. 118–134, Mar. 2014.
[34] N. J. Miller, "Removal of noise from a voice signal by synthesis," Utah Univ., Salt Lake City, UT, USA, Tech. Rep., 1973.
[35] A. V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," *IEEE Trans. Audio Electroacoust.*, vol. AE-16, no. 2, pp. 221–226, Jun. 1968.
[36] R. C. Maher, "An approach for the separation of voices in composite musical signals," Ph.D. dissertation, Univ. of Illinois at Urbana-Champaign, Champaign, IL, USA, 1989.
[37] A. L. Wang, "Instantaneous and frequency-warped techniques for auditory source separation," Ph.D. dissertation, Stanford Univ., Stanford, CA, USA, 1994.
[38] A. L. Wang, "Instantaneous and frequency-warped techniques for source separation and signal parametrization," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, New York, USA, Oct. 1995, pp. 47–50.
[39] Y. Meron and K. Hirose, "Separation of singing and piano sounds," in *Proc. 5th Int. Conf. Spoken Lang. Process.*, Sydney, NSW, Australia, Nov. 1998.
[40] T. F. Quatieri, "Shape invariant time-scale and pitch modification of speech," *IEEE Trans. Signal Process.*, vol. 40, no. 3, pp. 497–510, Mar. 1992.
[41] A. Ben-Shalom and S. Dubnov, "Optimal filtering of an instrument sound in a mixed recording given approximate pitch prior," in *Proc. Int. Comput. Music Conf.*, Miami, FL, USA, Nov. 2004.
[42] S. Shalev-Shwartz, S. Dubnov, N. Friedman, and Y. Singer, "Robust temporal and spectral modeling for query by melody," in *Proc. 25th Ann. Int. ACM SIGIR Conf. Res. Develop. Inf. Ret.*, Tampere, Finland, Aug. 2002, pp. 331–338.
[43] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*. Leiden, The Netherlands: Swets & Zeitlinger, 1997, pp. 91–122.
[44] B. V. Veen and K. M. Buckley, "Beamforming techniques for spatial filtering," in *Digital Signal Process. Handbook*. Boca Raton, FL, USA: CRC Press, 1997, pp. 1–22.
[45] 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, The Netherlands, Jul. 2005, pp. 562–565.
[46] Y.-G. Zhang and C.-S. Zhang, "Separation of music signals by harmonic structure modeling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1617–1624.
[47] E. Terhardt, "Calculating virtual pitch," *Hearing Res.*, vol. 1, no. 2, pp. 155–182, Mar. 1979.
[48] Y.-G. Zhang, C.-S. Zhang, and S. Wang, "Clustering in knowledge embedded space," in *Proc. Mach. Learn.* 2003, pp. 480–491.
[49] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Singer identification based on accompaniment sound reduction and reliable frame selection," in *Proc. 6th Int. Conf. Music Inf. Ret.*, London, U.K., Sep. 2005, pp. 329–336.
[50] 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.
[51] M. Goto, "A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Commun.*, vol. 43, no. 4, pp. 311–329, Sep. 2004.
[52] J. A. Moorer, "Signal processing aspects of computer music: A survey," *Proc. IEEE*, vol. 65, no. 8, pp. 1108–1137, Aug. 2005.
[53] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in *Proc. 7th Int. Conf. Music Inf. Ret.*, Victoria, BC, Canada, Oct. 2007, pp. 375–378.
[54] M. Ryynänen and A. Klapuri, "Transcription of the singing melody in polyphonic music," in *Proc. 7th Int. Conf. Music Inf. Ret.*, Victoria, BC, Canada, Oct. 2006.
[55] Z. Duan, Y.-F. Zhang, C.-S. 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.
[56] X. Rodet, "Musical sound signal analysis/synthesis: Sinusoidal+residual and elementary waveform models," in *Proc. IEEE Time, Freq. Time, Scale Workshop*, Coventry, U.K., Aug. 1997.
[57] J. O. Smith and X. Serra, "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *Proc. Int. Comput. Music Conf.*, Urbana, IL, USA, Aug. 1987, pp. 290–297.
[58] M. Slaney, D. Naar, and R. F. Lyon, "Auditory model inversion for sound separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Adelaide, SA, Australia, Apr. 1994, pp. II/77–II/80.
[59] M. Lagrange and G. Tzanetakis, "Sound source tracking and formation using normalized cuts," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Honolulu, HI, USA, Apr. 2007, pp. I-61–I-64.
[60] 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.
[61] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[62] 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, Aug. 2008, pp. 1417–1420.
[63] M. Ryynänen and A. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Comput. Music J.*, vol. 32, no. 3, pp. 72–86, Sep. 2008.
[64] Y. Ding and X. Qian, "Processing of musical tones using a combined quadratic polynomial-phase sinusoid and residual (QUASAR) signal model," *J. Audio Eng. Soc.*, vol. 45, no. 7/8, pp. 571–584, Jul. 1997.
[65] Y. Li and D. Wang, "Singing voice separation from monaural recordings," in *Proc. 7th Int. Conf. Music Inf. Ret*, Victoria, BC, Canada, Oct. 2006.
[66] 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.
[67] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, "Complex domain onset detection for musical signals," in *Proc. 6th Int. Conf. Digital Audio Effects*, London, U.K., Sep. 2003.
[68] Y. Li and D. Wang, "Detecting pitch of singing voice in polyphonic audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, USA, Mar. 2005, pp. 17–20.
[69] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 11, no. 3, pp. 229–241, May 2003.
[70] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2002.
[71] Y. Han and C. Raphael, "Desoloing monaural audio using mixture models," in *Proc. 7th Int. Conf. Music Inf. Ret.*, Victoria, BC, Canada, Oct. 2007, pp. 145–148.
[72] S. T. Roweis, "One microphone source separation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 793–799.
[73] C.-L. Hsu, J.-S. R. Jang, and T.-L. Tsai, "Separation of singing voice from music accompaniment with unvoiced sounds reconstruction for monaural recordings," in *Proc. AES 125th Conv.*, San Francisco, CA, USA, Oct. 2008, pp. 217–222.
[74] 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, Lang. Process.*, vol. 18, no. 2, pp. 310–319, Feb. 2010.
[75] K. Dressler, "Sinusoidal extraction using an efficient implementation of a multi-resolution FFT," in *Proc. 9th Int. Conf. Digit. Audio Effects*, Montreal, QC, Canada, Sep. 2006, pp. 247–252.
[76] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Atlanta, GA, USA, May 1996, pp. 629–632.
[77] C. Raphael and Y. Han, "A classifier-based approach to score-guided music audio source separation," *Comput. Music J.*, vol. 32, no. 1, pp. 51–59, 2008.
[78] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. London, U.K.: Chapman and Hall, 1984.

[79] E. Cano and C. Cheng, "Melody line detection and source separation in classical saxophone recordings," in *Proc. 12th Int. Conf. Digit. Audio Effects*, Como, Italy, Sep. 2009, pp. 478–483.
[80] S. Grollmisch, E. Cano, and C. Dittmar, "Songs2See: Learn to play by playing," in *Proc. Audio Eng. Soc. 41st Conf., Audio Games*, Feb. 2011.
[81] C. Dittmar, E. Cano, J. Abeßer, and S. Grollmisch, "Music information retrieval meets music education," in *Multimodal Music Processing*. Saarbrucken, Germany: Dagstuhl Publ., 2012, pp. 95–120.
[82] E. Cano, C. Dittmar, and G. Schuller, "Efficient implementation of a system for solo and accompaniment separation in polyphonic music," in *Proc. 20th Eur. Signal Process. Conf.*, Bucharest, Romania, Aug. 2012, pp. 285–289.
[83] K. Dressler, "Pitch estimation by the pair-wise evaluation of spectral peaks," in *Proc. 42nd Audio Eng. Soc. Conf. Semantic Audio*, Ilmenau, Germany, Jul. 2011.
[84] E. Cano, C. Dittmar, and G. Schuller, "Re-thinking sound separation: Prior information and additivity constraints in separation algorithms," in *Proc. 16th Int. Conf. Digit. Audio Effects*, Maynooth, Ireland, Sep. 2013.
[85] E. Cano, G. Schuller, and C. Dittmar, "Pitch-informed solo and accompaniment separation towards its use in music education applications," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 23, Sep. 2014.
[86] 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. 2012, pp. 2417–2421.
[87] R. Marxer, J. Janer, and J. Bonada, "Low-latency instrument separation in polyphonic audio using timbre models," in *Proc. 10th Int. Conf. Latent Var. Anal. Signal Separation*, Tel Aviv, Israel, Mar. 2012, pp. 314–321.
[88] A. Vaneph, E. McNeil, and F. Rigaud, "An automated source separation technology and its practical applications," in *Proc. 140th Audio Eng. Soc. Conv.*, Paris, France, May 2016.
[89] S. Leglaive, R. Hennequin, and R. Badeau, "Singing voice detection with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, Apr. 2015, pp. 121–125.
[90] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.
[91] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.
[92] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2003, pp. 177–180.
[93] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.
[94] C. Févotte, "Nonnegative matrix factorization with the Itakura–Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.
[95] P. Common, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, Apr. 1994.
[96] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *Proc. 6th Int. Conf. Music Inf. Ret.*, London, U.K., Sep. 2005, pp. 337–344.
[97] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.
[98] T. L. Nwe and Y. Wang, "Automatic detection of vocal segments in popular songs," in *Proc. 5th Int. Conf. Music Inf. Ret.*, Barcelona, Spain, Oct. 2004, pp. 138–145.
[99] M. A. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis," in *Proc. Int. Comput. Music Conf.*, Berlin, Germany, Sep. 2000.
[100] 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. 2008, pp. 243–246.
[101] A. Chanrungutai and C. A. Ratanamahatana, "Singing voice separation in mono-channel music," in *Proc. Int. Symp. Commun. Inf. Technol.*, Lao, China, Oct. 2008, pp. 256–261.
[102] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Soviet Math.*, vol. 4, pp. 1035–1038, 1963, pp. 1035–1038.
[103] R. Marxer and J. Janer, "A Tikhonov regularization method for spectrum decomposition in low latency audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 277–280.
[104] P.-K. Yang, C.-C. Hsu, and J.-T. Chien, "Bayesian singing-voice separation," in *Proc. 15th Int. Soc. Music Inf. Ret. Conf.*, Taipei, Taiwan, Oct. 2014, pp. 507–512.
[105] J.-T. Chien and P.-K. Yang, "Bayesian factorization and learning for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 1, pp. 185–195, Jan. 2015.
[106] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Comput. Intell. Neurosc.*, vol. 2009, no. 4, pp. 1–17, Jan. 2009.
[107] M. N. Schmidt, O. Winther, and L. K. Hansen, "Bayesian non-negative matrix factorization," in *Proc. 8th Int. Conf. Ind. Compon. Anal. Signal Separation*, Paraty, Brazil, Mar. 2009, pp. 540–547.
[108] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *Proc. 12th Int. Conf. Digit. Audio Effects*, Como, Italy, Sep. 2009.
[109] P. Smaragdis and G. J. Mysore, "Separation by "humming": User-guided sound extraction from monophonic mixtures," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, New York, USA, Oct. 2009, pp. 69–72.
[110] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. 7th Int. Conf. Ind. Compon. Anal. Signal Separation*, London, U.K., Sep. 2007, pp. 414–421.
[111] T. Nakamuray and H. Kameoka, "$L_p$-norm non-negative matrix factorization and its application to singing voice enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, Apr. 2015, pp. 2115–2119.
[112] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*. San Francisco, CA, USA: Academic, 1970.
[113] H. Kameoka, M. Goto, and S. Sagayama, "Selective amplifier of periodic and non-periodic components in concurrent audio signals with spectral control envelopes," Inf. Process. Soc. Japan, Tokyo, Japan, Tech. Rep. 2006-MUS-66, 2006.
[114] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 1–37, May 2011.
[115] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 57–60.
[116] P. Sprechmann, A. Bronstein, and G. Sapiro, "Real-time online singing voice separation from monaural recordings using robust low-rank modeling," in *Proc. 13th Int. Soc. Music Inform. Ret. Conf.*, Porto, Portugal, Oct. 2012, pp. 67–72.
[117] 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, Aug. 2010.
[118] B. Recht and C. Ré, "Parallel stochastic gradient algorithms for large-scale matrix completion," *Math. Program. Comput.*, vol. 5, no. 2, pp. 201–226, Jun. 2013.
[119] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, Jun. 2010, pp. 399–406.
[120] L. Zhang, Z. Chen, M. Zheng, and X. He, "Robust non-negative matrix factorization," *Frontiers Elect. Electron. Eng. China*, vol. 6, no. 2, pp. 192–200, Jun. 2011.
[121] I.-Y. Jeong and K. Lee, "Vocal separation using extended robust principal component analysis with Schatten $P/L_p$-norm and scale compression," in *Proc. Int. Workshop Mach. Learn. Signal Process.*, Reims, France, Nov. 2014.
[122] F. Nie, H. Wang, and H. Huang, "Joint Schatten $p$-norm and $l_p$-norm robust matrix completion for missing value recovery," *Knowl. Inf. Syst.*, vol. 42, no. 3, pp. 525–544, Mar. 2015.
[123] Y.-H. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," in *Proc. 14th Int. Soc. Music Inf. Ret. Conf.*, Curitiba, Brazil, Nov. 2013, pp. 427–432.
[124] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, Montreal, QC, Canada, Jun. 2009, pp. 689–696.
[125] T.-S. T. Chan and Y.-H. Yang, "Complex and quaternionic principal component pursuit and its application to audio separation," *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 287–291, Feb. 2016.
[126] G. Peeters, "Deriving musical structures from signal analysis for music audio summary generation: "sequence" and "state" approach," in *Proc. Int. Symp. Comput. Music Multidisciplinary Res.*, Montpellier, France, May 2003, pp. 143–166.

[127] R. B. Dannenberg and M. Goto, "Music structure analysis from acoustic signals," in *Handbook of Signal Processing in Acoustics*. New York, NY, USA: Springer, 2008, pp. 305–331.
[128] J. Paulus, M. Müller, and A. Klapuri, "Audio-based music structure analysis," in *Proc. 11th Int. Soc. Music Inf. Ret. Conf.*, Utrecht, The Netherlands, Aug. 2010, pp. 625–636.
[129] Z. Rafii and B. Pardo, "A simple music/voice separation system based on the extraction of the repeating musical structure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, May 2011, pp. 221–224.
[130] 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. 73–84, Jan. 2013.
[131] Z. Rafii, A. Liutkus, and B. Pardo, "REPET for background/foreground separation in audio," in *Blind Source Separation*. Berlin, Germany: Springer, 2014, pp. 395–411.
[132] J. Foote and S. Uchihashi, "The beat spectrum: A new approach to rhythm analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, Tokyo, Japan, Aug. 2001, pp. 881–884.
[133] P. Seetharaman, F. Pishdadian, and B. Pardo, "Music/voice separation using the 2d Fourier transform," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2017, pp. 36–40.
[134] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 53–56.
[135] Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix," in *Proc. 13th Int. Soc. Music Inf. Ret. Conf.*, Porto, Portugal, Oct. 2012, pp. 583–588.
[136] J. Foote, "Visualizing music and audio using self-similarity," in *Proc. 7th Assoc. Comput. Mach. Int. Conf. Multimedia*, Orlando, FL, USA, Oct. 1999, pp. 77–80.
[137] Z. Rafii and B. Pardo, "Online REPET-SIM for real-time speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 848–852.
[138] Z. Rafii, A. Liutkus, and B. Pardo, "A simple user interface system for recovering patterns repeating in time and frequency in mixtures of sounds," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, Apr. 2015, pp. 271–275.
[139] D. FitzGerald, "Vocal separation using nearest neighbours and median filtering," in *Proc. 23rd IET Irish Signals Syst. Conf.*, Maynooth, Ireland, Jun. 2012.
[140] 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*, Villers-les-Nancy, France, May 2014.
[141] A. Liutkus, D. FitzGerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4298–4310, Aug. 2014.
[142] A. Liutkus, D. FitzGerald, and Z. Rafii, "Scalable audio separation with light kernel additive modelling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, Apr. 2015, pp. 76–80.
[143] T. Prätzlich, R. Bittner, A. Liutkus, and M. Müller, "Kernel additive modeling for interference reduction in multi-channel music recordings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, Aug. 2015, pp. 584–588.
[144] D. F. Yela, S. Ewert, D. FitzGerald, and M. Sandler, "Interference reduction in music recordings combining kernel additive modelling and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 51–55.
[145] M. Moussallam, G. Richard, and L. Daudet, "Audio source separation informed by redundancy with greedy multiscale decompositions," in *Proc. 20th Eur. Signal Process. Conf.*, Bucharest, Romania, Aug. 2012, pp. 2644–2648.
[146] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
[147] H. Deif, D. FitzGerald, W. Wang, and L. Gan, "Separation of vocals from monaural music recordings using diagonal median filters and practical time-frequency parameters," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol.*, Abu Dhabi, United Arab Emirates, Dec. 2015, pp. 163–167.
[148] 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, Jan. 2010.
[149] J.-Y. Lee and H.-G. Kim, "Music and voice separation using log-spectral amplitude estimator based on kernel spectrogram models backfitting," *J. Acoust. Soc. Korea*, vol. 34, no. 3, pp. 227–233, 2015.
[150] J.-Y. Lee, H.-S. Cho, and H.-G. Kim, "Vocal separation from monaural music using adaptive auditory filtering based on kernel back-fitting," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015, pp. 3317–3320.
[151] H.-S. Cho, J.-Y. Lee, and H.-G. Kim, "Singing voice separation from monaural music based on kernel back-fitting using beta-order spectral amplitude estimation," in *Proc. 16th Int. Soc. Music Inf. Ret. Conf.*, Málaga, Spain, Oct. 2015, pp. 639–644.
[152] H.-G. Kim and J. Y. Kim, "Music/voice separation based on kernel back-fitting using weighted $\beta$-order MMSE estimation," *ETRI J.*, vol. 38, no. 3, pp. 510–517, Jun. 2016.
[153] E. Plourde and B. Champagne, "Auditory-based spectral amplitude estimators for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1614–1623, Nov. 2008.
[154] 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, 2007.
[155] P. Smaragdis and B. Raj, "Shift-invariant probabilistic latent component analysis," Mitubishi Elect. Res. Lab., Cambridge, MA, USA, Tech. Rep. TR2007-009, 2006.
[156] B. Raj and P. Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2005, pp. 17–20.
[157] J. Han and C.-W. Chen, "Improving melody extraction using probabilistic latent component analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, May 2011, pp. 33–36.
[158] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. Inst. Phonetic Sci., Univ. Amsterdam, Proc. 17*, 1993, pp. 97–110.
[159] E. Gómez, F. J. C. nadas Quesada, J. Salamon, J. Bonada, P. V. Candea, and P. C. nas Molero, "Predominant fundamental frequency estimation vs singing voice separation for the automatic transcription of accompanied flamenco singing," in *Proc. 13th Int. Soc. Music Inf. Ret. Conf.*, Porto, Portugal, Aug. 2012.
[160] N. Ono, K. Miyamoto, J. L. Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. 16th Eur. Signal Process. Conf.*, Lausanne, Switzerland, Aug. 2008.
[161] 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. 2014.
[162] T.-S. Chan *et al.*, "Vocal activity informed singing voice separation with the iKala dataset," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, Apr. 2015, pp. 718–722.
[163] I.-Y. Jeong and K. Lee, "Singing voice separation using RPCA with weighted $l_1$-norm," in *Proc. 13th Int. Conf. Latent Var. Anal. Signal Separation*, Grenoble, France, Feb. 2017, pp. 553–562.
[164] 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 Tuts. Res. Workshop Stat. Perceptual Audition*, Brisbane, Australia, Sep. 2008.
[165] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, May 2011.
[166] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Proc. 7th Int. Conf. Music Inf. Ret.*, Victoria, BC, Canada, Oct. 2006, pp. 216–221.
[167] C.-L. Hsu, L.-Y. Chen, J.-S. R. Jang, and H.-J. Li, "Singing pitch extraction from monaural polyphonic songs by contextual audio modeling and singing harmonic enhancement," in *Proc. 10th Int. Soc. Music Inf. Ret. Conf.*, Kyoto, Japan, Oct. 2009, pp. 201–206.
[168] Z. Rafii, Z. Duan, and B. Pardo, "Combining rhythm-based and pitch-based methods for background and melody separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1884–1893, Sep. 2014.
[169] 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.
[170] S. Venkataramani, N. Nayak, P. Rao, and R. Velmurugan, "Vocal separation using singer-vowel priors obtained from polyphonic audio," in *Proc. 15th Int. Soc. Music Inf. Ret. Conf.*, Taipei, Taiwan, Oct. 2014, pp. 283–288.

[171] V. Rao and P. Rao, "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2145–2154, Nov. 2010.
[172] V. Rao, C. Gupta, and P. Rao, "Context-aware features for singing voice detection in polyphonic music," in *Proc. Int. Workshop Adapt. Multimedia Ret.*, Barcelona, Spain, Jul. 2011, pp. 43–57.
[173] M. Kim, J. Yoo, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for spectral and temporal drum source separation," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1192–1204, Oct. 2011.
[174] L. Zhang, Z. Chen, M. Zheng, and X. He, "Nonnegative matrix and tensor factorizations: An algorithmic perspective," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 54–65, May 2014.
[175] Y. Ikemiya, K. Yoshii, and K. Itoyama, "Singing voice analysis and editing based on mutually dependent F0 estimation and source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, Apr. 2015, pp. 574–578.
[176] Y. Ikemiya, K. Itoyama, and K. Yoshii, "Singing voice separation and vocal F0 estimation based on mutual combination of robust principal component analysis and subharmonic summation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2084–2095, Nov. 2016.
[177] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Amer.*, vol. 83, no. 1, pp. 257–264, Jan. 1988.
[178] A. Dobashi, Y. Ikemiya, K. Itoyama, and K. Yoshii, "A music performance assistance system based on vocal, harmonic, and percussive source separation and content visualization for music audio signals," in *Proc. 12th Sound Music Comput. Conf.*, Maynooth, Ireland, Jul. 2015.
[179] Y. Hu and G. Liu, "Separation of singing voice using nonnegative matrix partial co-factorization for singer identification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 643–653, Apr. 2015.
[180] J. Yoo, M. Kim, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for drum source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 1942–945.
[181] P. Boersma, "PRAAT, a system for doing phonetics by computer," *Glot Int.*, vol. 5, no. 9/10, pp. 341–347, Dec. 2001.
[182] Y. Li, J. Woodruff, and D. Wang, "Monaural musical sound separation based on pitch and common amplitude modulation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1361–1371, Sep. 2009.
[183] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Commun.*, vol. 43, no. 4, pp. 275–296, Sep. 2004.
[184] Y. Hu and G. Liu, "Monaural singing voice separation by non-negative matrix partial co-factorization with temporal continuity and sparsity criteria," in *Proc. 12th Int. Conf. Intell. Comput.*, Lanzhou, China, Aug. 2016, pp. 33–43.
[185] X. Zhang, W. Li, and B. Zhu, "Latent time-frequency component analysis: A novel pitch-based approach for singing voice separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, Apr. 2015, pp. 131–135.
[186] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.
[187] B. Zhu, W. Li, and L. Li, "Towards solving the bottleneck of pitch-based singing voice separation," in *Proc. 23rd Assoc. Comput. Mach. Int. Conf. Multimedia*, Brisbane, QLD, Australia, Oct. 2015, pp. 511–520.
[188] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, USA, Apr. 2008, pp. 169–172.
[189] J.-L. Durrieu, G. Richard, and B. David, "An iterative approach to monaural musical mixture de-soloing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 105–108.
[190] 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.
[191] 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.
[192] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Amer.*, vol. 87, no. 2, pp. 820–857, Feb. 1990.
[193] L. Benaroya, L. Mcdonagh, F. Bimbot, and R. Gribonval, "Non negative sparse representation for Wiener based source separation with a single sensor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, China, Apr. 2003, pp. 613–616.
[194] I. S. Dhillon and S. Sra, "Generalized nonnegative matrix approximations with Bregman divergences," in *Proc. Adv. Neural Inf. Process. Syst 18*, 2005, pp. 283–290.
[195] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 191–199, Jan. 2006.
[196] J.-L. Durrieu and J.-P. Thiran, "Musical audio source separation based on user-selected F0 track," in *Proc. 10th Int. Conf. Latent Var. Anal. Signal Separation*, Tel Aviv, Israel, Mar. 2012, pp. 438–445.
[197] B. Fuentes, R. Badeau, and G. Richard, "Blind harmonic adaptive decomposition applied to supervised source separation," in *Proc. IEEE 20th Eur. Signal Process. Conf.*, 2012, pp. 2654–2658.
[198] J. C. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, Jan. 1991.
[199] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant Q transform," *J. Acoust. Soc. Amer.*, vol. 92, no. 5, pp. 2698–2701, Nov. 1992.
[200] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox," in *Proc. 7th Sound Music Comput. Conf.*, Barcelona, Spain, Jul. 2010.
[201] 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.
[202] C. Joder and B. Schuller, "Score-informed leading voice separation from monaural audio," in *Proc. 13th Int. Soc. Music Inf. Ret. Conf.*, Porto, Portugal, Oct. 2012, pp. 277–282.
[203] C. Joder, S. Essid, and G. Richard, "A conditional random field framework for robust and scalable audio-to-score matching," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2385–2397, Nov. 2011.
[204] R. Zhao, S. Lee, D.-Y. Huang, and M. Dong, "Soft constrained leading voice separation with music score guidance," in *Proc. 9th Int. Symp. Chin. Spoken Lang.*, Singapore, Sep. 2014, pp. 565–569..
[205] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," in *Proc. 17th Eur. Signal Process. Conf.*, Glasgow, U.K., Aug. 2009, pp. 15–19.
[206] J. Janer and R. Marxer, "Separation of unvoiced fricatives in singing voice mixtures with semi-supervised NMF," in *Proc. 16th Int. Conf. Digit. Audio Effects*, Maynooth, Ireland, Sep. 2013.
[207] J. Janer, R. Marxer, and K. Arimoto, "Combining a harmonic-based NMF decomposition with transient analysis for instantaneous percussion separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 281–284.
[208] R. Marxer and J. Janer, "Modelling and separation of singing voice breathiness in polyphonic mixtures," in *Proc. 16th Int. Conf. Digit. Audio Effects*, Maynooth, Ireland, Sep. 2013.
[209] G. Degottex, A. Roebel, and X. Rodet, "Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, May 2011, pp. 5128–5131.
[210] A. Ozerov, E. Vincent, and F. Bimbot, "A general modular framework for audio source separation," in *Proc. 9th Int. Conf. Latent Var. Anal. Signal Separation*, St. Malo, France, Sep. 2010, pp. 33–40.
[211] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1118–1133, May 2012.
[212] Y. Salaün *et al.*, "The flexible audio source separation toolbox version 2.0," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 2014.
[213] R. Hennequin and F. Rigaud, "Long-term reverberation modeling for under-determined audio source separation with application to vocal melody extraction," in *Proc. 17th Int. Soc. Music Inf. Ret. Conf.*, New York City, NY, USA, Aug. 2016.
[214] R. Singh, B. Raj, and P. Smaragdis, "Latent-variable decomposition based dereverberation of monaural and multi-channel signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, Mar. 2010, pp. 1914–1917.
[215] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *Proc. 9th Int. Conf. Music Inf. Ret.*, Philadelphia, PA, USA, Sep. 2008, pp. 139–144.
[216] D. FitzGerald, "Harmonic/percussive separation using median filtering," in *Proc. 13th Int. Conf. Digital Audio Effects*, Graz, Austria, Sep. 2010.

[217] Y.-H. Yang, "On sparse and low-rank matrix decomposition for singing voice separation," in *Proc. 20th ACM Int. Conf. Multimedia*, Nara, Japan, Oct. 2012, pp. 757–760.
[218] I.-Y. Jeong and K. Lee, "Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints," *IEEE Signal Process. Lett.*, vol. 21, no. 10, pp. 1197–1200, Jun. 2014.
[219] E. Ochiai, T. Fujisawa, and M. Ikehara, "Vocal separation by constrained non-negative matrix factorization," in *Proc. Asia, Pac. Signal Inf. Process. Assoc. Annu. Summit Conf.*, Hong Kong, China, Dec. 2015, pp. 480–483.
[220] T. Watanabe, T. Fujisawa, and M. Ikehara, "Vocal separation using improved robust principal component analysis and post-processing," in *Proc. IEEE 59th Int. Midwest Symp. Circuits Syst.*, Abu Dhabi, United Arab Emirates, Oct. 2016.
[221] H. Raguet, J. Fadili, and G. Peyré, "A generalized forward-backward splitting," *SIAM J. Imag. Sci.*, vol. 6, no. 3, pp. 1199–1226, Jul. 2013.
[222] A. Hayashi, H. Kameoka, T. Matsubayashi, and H. Sawada, "Non-negative periodic component analysis for music source separation," in *Proc. Asia, Pac. Signal Inf. Process. Assoc. Annu. Summit Conf.*, Jeju, South Korea, Dec. 2016.
[223] D. FitzGerald, M. Cranitch, and E. Coyle, "Using tensor factorisation models to separate drums from polyphonic music," in *Proc. 12th Int. Conf. Digit. Audio Effects*, Como, Italy, Sep. 2009.
[224] 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.
[225] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, Mar. 2010, pp. 425–428.
[226] H. Tachibana, N. Ono, and S. Sagayama, "A real-time audio-to-audio Karaoke generation system for monaural recordings based on singing voice suppression and key conversion techniques," *J. Inf. Process.*, vol. 24, no. 3, pp. 470–482, May 2016.
[227] N. Ono *et al.*, "Harmonic and percussive sound separation and its application to MIR-related tasks," in *Proc.Adv. Music Inf. Ret.*, 2010, pp. 213–236.
[228] H. Tachibana, H. Kameoka, N. Ono, and S. Sagayama, "Comparative evaluations of multiple harmonic/percussive sound separation techniques based on anisotropic smoothness of spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 465–468.
[229] H. Deif, W. Wang, L. Gan, and S. Alhashmi, "A local discontinuity based approach for monaural singing voice separation from accompanying music with multi-stage non-negative matrix factorization," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Orlando, FL, USA, Dec. 2015, pp. 93–97.
[230] 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. 2096–2107, Oct. 2013.
[231] J. Driedger and M. Müller, "Extracting singing voice from music recordings by cascading audio decomposition techniques," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, Apr. 2015, pp. 126–130.
[232] J. Driedger, M. Müller, and S. Disch, "Extending harmonic-percussive separation of audio signals," in *Proc. 15th Int. Soc. Music Inf. Retrieval Conf.*, Taipei, Taiwan, Oct. 2014, pp. 611–616.
[233] R. Talmon, I. Cohen, and S. Gannot, "Transient noise reduction using nonlocal diffusion filters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1584–1599, Aug. 2011.
[234] 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.
[235] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.
[236] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1*. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362.
[237] N. J. Bryan and G. J. Mysore, "Interactive user-feedback for sound source separation," in *Proc. Int. Conf. Intell. User, Interfaces, Workshop Interact. Mach. Learn.*, Santa Monica, CA, USA, Mar. 2013.
[238] N. J. Bryan and G. J. Mysore, "An efficient posterior regularized latent variable model for interactive sound source separation," in *Proc. 30th Int. Conf. Mach. Learn.*, Atlanta, GA, USA, Jun. 2013, pp. III-208–III-216.
[239] N. J. Bryan and G. J. Mysore, "Interactive refinement of supervised and semi-supervised sound source separation estimates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 883–887.
[240] K. Ganchev, J. Ao Graça, J. Gillenwater, and B. Taskar, "Posterior regularization for structured latent variable models," *J. Mach. Learn. Res.*, vol. 11, pp. 2001–2049, Mar. 2010.
[241] A. Ozerov, N. Duong, and L. Chevallier, "Weighted nonnegative tensor factorization: on monotonicity of multiplicative update rules and application to user-guided audio source separation," Technicolor, Tech. Rep., 2013.
[242] X. Jaureguiberry, G. Richard, P. Leveau, R. Hennequin, and E. Vincent, "Introducing a simple fusion framework for audio source separation," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Southampton, U.K., Sep. 2013.
[243] X. Jaureguiberry, E. Vincent, and G. Richard, "Variational Bayesian model averaging for audio source separation," in *Proc. IEEE Workshop Stat. Signal Process. Workshop*, Gold Coast, VIC, Australia, Jun. 2014, pp. 33–36.
[244] X. Jaureguiberry, E. Vincent, and G. Richard, "Fusion methods for speech enhancement and audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 7, pp. 1266–1279, Jul. 2016.
[245] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: A tutorial," *Stat. Sci.*, vol. 14, no. 4, pp. 382–417, Nov. 1999.
[246] M. McVicar, R. Santos-Rodriguez, and T. De Bie, "Learning to separate vocals from polyphonic mixtures via ensemble methods and structured output prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, Mar. 2016, pp. 450–454.
[247] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Los Angeles, CA, USA, Nov. 1990, pp. 14–19.
[248] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *Proc. 15th Int. Soc. Music Inf. Ret. Conf.*, Taipei, Taiwan, Oct. 2014.
[249] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher, "Block-coordinate Frank-Wolfe optimization for structural SVMs," in *Proc. 30th Int. Conf. Mach. Learn.*, Atlanta, GA, USA, Jun. 2013, pp. I-53–I-61.
[250] E. Manilow, P. Seetharaman, F. Pishdadian, and B. Pardo, "Predicting algorithm efficacy for adaptive, multi-cue source separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2017, pp. 274–278.
[251] G. Wolf, S. Mallat, and S. Shamma, "Audio source separation with time-frequency velocities," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Reims, France, Sep. 2014.
[252] G. Wolf, S. Mallat, and S. Shamma, "Rigid motion model for audio source separation," *IEEE Trans. Signal Process.*, vol. 64, no. 7, pp. 1822–1831, Apr. 2016.
[253] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128, Aug. 2014.
[254] C. P. Bernard, "Discrete wavelet analysis for fast optic flow computation," *Appl. Comput. Harmon. Anal.*, vol. 11, no. 1, pp. 32–63, Jul. 2001.
[255] F. Yen, Y.-J. Luo, and T.-S. Chi, "Singing voice separation using spectro-temporal modulation features," in *Proc. 15th Int. Soc. Music Inf. Ret. Conf.*, Taipei, Taiwan, Oct. 2014, pp. 617–622.
[256] F. Yen, M.-C. Huang, and T.-S. Chi, "A two-stage singing voice separation algorithm using spectro-temporal modulation features," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015, pp. 3321–3324.
[257] T. Chi, P. Rub, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, no. 2, pp. 887–906, Aug. 2005.
[258] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 106, no. 5, pp. 2719–2732, Nov. 1999.
[259] T.-S. T. Chan and Y.-H. Yang, "Informed group-sparse representation for singing voice separation," *IEEE Signal Process. Lett.*, vol. 24, no. 2, pp. 156–160, Feb. 2017.
[260] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Stat. Soc. Ser. B*, vol. 68, no. 1, pp. 49–67, Dec. 2006.
[261] S. Ma, "Alternating proximal gradient method for convex minimization," *J. Sci. Comput.*, vol. 68, no. 2, p. 546572, Aug. 2016.
[262] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2007.

[263] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.
[264] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," 1993.
[265] N. Sturmel *et al.*, "Linear mixing models for active listening of music productions in realistic studio conditions," in *Proc. 132nd AES Convention*, Budapest, Hungary, Apr. 2012.
[266] M. Vinyes, "MTG MASS database," 2008. [Online]. Available: http://www.mtg.upf.edu/static/mass/resources
[267] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. 8th Int. Conf. Ind. Compon. Anal. Signal Separation*, Paraty, Brazil, Mar. 2009, pp. 734–741.
[268] S. Araki *et al.*, "The 2010 signal separation evaluation campaign (SiSEC2010): - Audio source separation -," in *Proc. 9th Int. Conf. Latent Var. Anal. Signal Separation*, St. Malo, France, Sep. 2010, pp. 114–122.
[269] S. Araki et al.,"The 2011 signal separation evaluation campaign (SiSEC2011): - Audio source separation -," in *Proc. 10th Int. Conf. Latent Var. Anal. Signal Separation*, 2012, pp. 414–422.
[270] E. Vincent *et al.*, "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Signal Process.*, vol. 92, no. 8, pp. 1928–1936, Aug. 2012.
[271] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *Proc. 12th Int. Conf. Latent Var. Anal. Signal Separation*, Liberec, Czech Republic, Aug. 2015, pp. 387–395.
[272] A. Liutkus *et al.*, "The 2016 signal separation evaluation campaign," in *Proc. 13th Int. Conf. Latent Var. Anal. Signal Separation*, Grenoble, France, Feb. 2017, pp. 323–332.
[273] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 59, no. 7, pp. 3155–3167, Feb. 2011.
[274] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive MIR research," in *Proc. 15th Int. Soc. Music Inf. Ret. Conf.*, Taipei, Taiwan, Oct. 2014.
[275] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "Musdb18, a dataset for audio source separation," Dec. 2017. [Online]. Available: https://sigsep.github.io/musdb
[276] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2005, pp. 90–93.
[277] W.-H. Tsai, D. Rogers, and H.-M. Wang, "Blind clustering of popular music recordings based on singer voice characteristics," *Comput. Music J.*, vol. 28, no. 3, pp. 68–78, 2004.
[278] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.
[279] E. Vincent, M. Jafari, S. Abdallah, M. Plumbley, and M. Davies, "Probabilistic modeling paradigms for audio source separation," *Mach. Audition, Principles, Algorithms Syst.*, IGI Global, 2010, pp. 162–185.
[280] 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. Ret. Conf.*, Curitiba, PR, Brazil, Nov. 2013.
[281] N. Boulanger-Lewandowski, G. J. Mysore, and M. Hoffman, "Exploiting long-term temporal dependencies in NMF using recurrent neural networks with application to source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 2014, pp. 6969–6973.
[282] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in *Proc. 9th Int. Conf. Latent Var. Anal. Signal Separation*, St. Malo, France, Sep. 2010, pp. 140–148.
[283] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using Bayesian wavenet," in *Proc. Interspeech 2017*, 2017, pp. 2013–2017.
[284] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, no. 3/4, pp. 197–387, Jun. 2014.
[285] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[286] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[287] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.*, vol. 22, no. 3, pp. 400–407, Sep. 1951.
[288] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986.
[289] M. Hermans and B. Schrauwen, "Training and analysing deep recurrent neural networks," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2013, pp. 190–198.
[290] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," in *Proc. Int. Conf. Learn. Representations*, Banff, AB, Canada, Apr. 2014.
[291] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.
[292] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 2014, pp. 1562–1566.
[293] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, Apr. 2015, pp. 2135–2139.
[294] S. Uhlich *et al.*, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 261–265.
[295] A. J. R. Simpson, G. Roma, and M. D. Plumbley, "Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network," in *Proc. 12th Int. Conf. Latent Var. Anal. Signal Separation*, Liberec, Czech Republic, Aug. 2015, pp. 429–436.
[296] J. Schlüter, "Learning to pinpoint singing voice from weakly labeled examples," in *Proc. 17th Int. Soc. Music Inf. Ret. Conf.*, New York City, NY, USA, Aug. 2016, pp. 44–50.
[297] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *Proc. 13th Int. Conf. Latent Var. Anal. Signal Separation*, Grenoble, France, Feb. 2017, pp. 258–266.
[298] S. I. Mimilakis, E. Cano, J. Abeßer, and G. Schuller, "New sonorities for jazz recordings: Separation and mixing using deep neural networks," in *Proc. 2nd AES Workshop Intell. Music Prod.*, London, U.K., Sep. 2016.
[299] S. I. Mimilakis, K. Drossos, T. Virtanen, and G. Schuller, "A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process.*, Tokyo, Japan, Sep. 2017.
[300] S. I. Mimilakis, K. Drossos, J. ao F. Santos, G. Schuller, T. Virtanen, and Y. Bengio, "Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask," in *Proc. IEEE 27th Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018.
[301] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. 18th Int. Soc. Music Inf. Ret. Conf.*, Suzhou, China, Oct. 2017, pp. 745–751.
[302] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2017, pp. 21–25.
[303] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, Mar. 2016, pp. 31–35.
[304] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multispeaker separation using deep clustering," in *Proc. Interspeech*, San Francisco, CA, USA, 2016, pp. 545–549.
[305] Y. Luo, Z. Chen, J. R. Hershey, J. L. Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New-Orleans, LA, USA, Mar. 2017, pp. 61–65.
[306] M. Kim and P. Smaragdis, "Adaptive denoising autoencoders: A fine-tuning scheme to learn from test mixtures," in *Proc. 12th Int. Conf. Latent Var. Anal. Signal Separation*, Liberec, Czech Republic, Aug. 2015, pp. 100–107.
[307] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.
[308] E. M. Grais, G. Roma, A. J. R. Simpson, and M. D. Plumbley, "Single channel audio source separation using deep neural network ensembles," in *Proc. 140th Audio Eng. Soc. Conv.*, Paris, France, May 2016.

[309] E. M. Grais, G. Roma, A. J. R. Simpson, and M. D. Plumbley, "Combining mask estimates for single channel audio source separation using deep neural networks," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 3339–3343.

[310] E. M. Grais, G. Roma, A. J. R. Simpson, and M. D. Plumbley, "Discriminative enhancement for single channel audio source separation using deep neural networks," in *Proc. 13th Int. Conf. Latent Var. Anal. Signal Separation*, Grenoble, France, Feb. 2017, pp. 236–246.

[311] E. M. Grais, G. Roma, A. J. R. Simpson, and M. D. Plumbley, "Two-stage single-channel audio source separation using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 9, pp. 1773–1783, Sep. 2017.

[312] S. Nie *et al.*, "Joint optimization of recurrent networks exploiting source auto-regression for source separation," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015.

[313] J. Sebastian and H. A. Murthy, "Group delay based music source separation using deep recurrent neural networks," in *Proc. Int. Conf. Signal Process. Commun.*, Bangalore, India, Jun. 2016.

[314] B. Yegnanarayana, H. A. Murthy, and V. R. Ramachandran, "Processing of noisy speech using modified group delay functions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toronto, ON, Canada, Apr. 1991, pp. 945–948.

[315] Z.-C. Fan, J.-S. R. Jang, and C.-L. Lu, "Singing voice separation and pitch extraction from monaural polyphonic audio music via DNN and adaptive pitch tracking," in *Proc. IEEE 2nd Int. Conf. Multimedia Big Data*, Taipei, Taiwan, Apr. 2016, pp. 178–185.

[316] C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2003, pp. 55–58.

[317] C. Avendano and J.-M. Jot, "Frequency domain techniques for stereo to multichannel upmix," in *Proc. 22nd Int. Conf. Audio Eng. Soc.*, Espoo, Finland, Jun. 2002.

[318] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discrimination and resynthesis," in *Proc. 7th Int. Conf. Digit. Audio Effects*, Naples, Italy, Oct. 2004.

[319] M. Vinyes, J. Bonada, and A. Loscos, "Demixing commercial music productions via human-assisted time-frequency masking," in *Proc. 120th Audio Eng. Soc. Conv.*, Paris, France, May 2006.

[320] M. Cobos and J. J. López, "Stereo audio source separation based on time-frequency masking and multilevel thresholding," *Digit. Signal Process.*, vol. 18, no. 6, pp. 960–976, Nov. 2008.

[321] Özgür Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[322] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[323] S. Sofianos, A. Ariyaeeinia, and R. Polfreman, "Towards effective singing voice extraction from stereophonic recordings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, Mar. 2010, pp. 233–236.

[324] S. Sofianos, A. Ariyaeeinia, and R. Polfreman, "Singing voice separation based on non-vocal independent component subtraction and amplitude discrimination," in *Proc. 13th Int. Conf. Digit. Audio Effects*, Graz, Austria, Sep. 2010, pp. 221–224.

[325] S. Sofianos, A. Ariyaeeinia, R. Polfreman, and R. Sotudeh, "H-semantics: A hybrid approach to singing voice separation," *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 831–841, Oct. 2012.

[326] M. Kim, S. Beack, K. Choi, and K. Kang, "Gaussian mixture model for singing voice separation from stereophonic music," in *Proc. Audio Eng. Soc. 43rd Conf.*, Pohang, South Korea, Sep. 2011.

[327] M. Cobos and J. J. López, "Singing voice separation combining panning information and pitch tracking," in *Proc. Audio Eng. Soc. 124th Conv.*, Amsterdam, The Netherlands, May 2008, pp. 786–794.

[328] D. FitzGerald, "Stereo vocal extraction using ADRess and nearest neighbours median filtering," in *Proc. 16th Int. Conf. Digital Audio Effects*, Maynooth, Ireland, Jan. 2013.

[329] D. FitzGerald and R. Jaiswal, "Improved stereo instrumental track recovery using median nearest-neighbour inpainting," in *Proc. 24th IET Irish Signals Syst. Conf.*, Letterkenny, Ireland, Jun. 2013.

[330] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 922–932, Mar. 2012.

[331] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures with application to blind audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 3137–3140.

[332] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[333] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, May 2011, pp. 257–260.

[334] A. Liutkus, R. Badeau, and G. Richard, "Informed source separation using latent components," in *Proc. 9th Int. Conf. Latent Var. Anal. Signal Separation*, St. Malo, France, Sep. 2010, pp. 498–505.

[335] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: Statistical insights and towards self-clustering of the spatial cues," in *Proc. 7th Int. Symp. Comput. Music Model. Ret.*, Málaga, Spain, Jun. 2010, pp. 102–115.

[336] A. Ozerov, N. Duong, and L. Chevallier, "On monotonicity of multiplicative update rules for weighted nonnegative tensor factorization," in *Proc. Int. Symp. Nonlinear Theory Appl.*, Luzern, Switzerland, Sep. 2014, pp. 40–43.

[337] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "New formulations and efficient algorithms for multichannel NMF," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, New York, USA, Oct. 2011, pp. 153–156.

[338] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 261–264.

[339] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 971–982, May 2013.

[340] S. Sivasankaran *et al.*, "Robust ASR using neural network based speech enhancement and feature simulation," in *Proc. IEEE Automat. Speech Recog. Understanding Workshop*, Scottsdale, AZ, USA, Dec. 2015, pp. 482–489.

[341] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, Sep. 2016.

[342] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," Inria, France, Tech. Rep. RR-8740, 2015.

[343] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel music separation with deep neural networks," in *Proc. 24th Eur. Signal Process. Conf.*, Budapest, Hungary, Aug. 2016, pp. 1748–1752.

[344] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.

[345] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Informed source separation: Source coding meets source separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2011, pp. 257–260.

[346] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Berlin, Germany: Springer, 2013.

[347] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Salt Lake City, UT, USA, May 2001, pp. 749–752.

[348] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[349] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Automat. Speech Recognit. Understanding*, Scottsdale, AZ, USA, Dec. 2015, pp. 504–511.

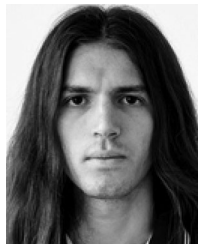
[350] I. Recommendation, "Bs. 1534-1. method for the subjective assessment of intermediate sound quality (MUSHRA)," *Int. Telecommun. Union, Geneva*, 2001.

[351] E. Vincent, M. Jafari, and M. Plumbley, "Preliminary guidelines for subjective evaluation of audio source separation algorithms," in *Proc. ICA Res. Netw. Int. Workshop*, Southampton, U.K., Sep. 2006.
[352] E. Cano, C. Dittmar, and G. Schuller, "Influence of phase, magnitude and location of harmonic components in the perceived quality of extracted solo signals," in *Proc. Audio Eng. Soc. 42nd Conf. Semantic Audio*, Ilmenau, Germany, Jul. 2011, pp. 247–252.
[353] C. Févotte, R. Gribonval, and E. Vinvent, "BSS_EVAL toolbox user guide - revision 2.0," IRISA, Rennes, France, Tech. Rep. 1706, 2005.
[354] 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.
[355] 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. Latent Var. Anal. Signal Separation*, London, U.K., Sep. 2007, pp. 454–461.
[356] B. Fox and B. Pardo, "Towards a model of perceived quality of blind audio source separation," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Beijing, China, Jul. 2007, pp. 1898–1901.
[357] J. Kornycky, B. Gunel, and A. Kondoz, "Comparison of subjective and objective evaluation methods for audio source separation," *J. Acoust. Soc. Amer.*, vol. 4, no. 1, 2008, Art. no. 050001.
[358] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Multi-criteria subjective and objective evaluation of audio source separation," in *Proc. 38th Int. Audio Eng. Soc. Conf.*, Pitea, Sweden, Jun. 2010.
[359] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.
[360] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *Proc. 10th Int. Conf. Latent Var. Anal. Signal Separation*, Tel Aviv, Israel, Mar. 2012, pp. 430–437.
[361] M. Cartwright, B. Pardo, G. J. Mysore, and M. Hoffman, "Fast and easy crowdsourced perceptual audio evaluation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, Mar. 2016, pp. 619–623.
[362] U. Gupta, E. Moore, and A. Lerch, "On the perceptual relevance of objective source separation measures for singing voice separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2005.
[363] F.-R. Stöter, A. Liutkus, R. Badeau, B. Edler, and P. Magron, "Common fate model for unison source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, Mar. 2016, pp. 126–130.
[364] G. Roma, E. M. Grais, A. J. Simpson, I. Sobieraj, and M. D. Plumbley, "Untwist: A new toolbox for audio source separation," in *Proc. 17th Int. Soc. Music Inf. Ret. Conf.*, New York City, NY, USA, Aug. 2016.

**Zafar Rafii** (M'14) received the M.S. degree in electrical engineering from both Ecole Nationale Supérieure de l'Electronique et de ses Applications, France, and Illinois Institute of Technology, Chicago, IL, USA, in 2006, and the Ph.D. degree in electrical engineering and computer science from the Northwestern University, Evanston, IL, USA, in 2014. He is currently a Senior Research Engineer with Gracenote, Emeryville, CA, USA. He was a Research Engineer with Audionamix, Paris, France. His research interests are include audio analysis, somewhere between signal processing, machine learning, and cognitive science, with a predilection for source separation and audio identification in music.

**Antoine Liutkus** (M'12) received the State Engineering degree in 2005 from Télécom ParisTech, Paris, France, where he received the Ph.D. degree in electrical engineering in 2012, and the M.Sc. degree in acoustics, computer science and signal processing applied to music from the Université Pierre et Marie Curie (Paris VI), Paris, France, in 2005. From 2007 to 2010, he was a Research Engineer on source separationwith the Audionamix, Paris, France. He is currently a Researcher with Inria, France. His research interests include audio source separation and machine learning.

**Fabian-Robert Stöter** (S'12) received the Diploma degree in electrical engineering from the Leibniz Universität Hannover, Hannover, Germany, in 2012. He is currently working toward the Ph.D. degree in audio signal processing in the research group of B. Edler with the International Audio Laboratories Erlangen, Erlangen, Germany. He is currently a Researcher with Inria, France. His research interests include supervised and unsupervised methods for audio source separation and signal analysis of highly overlapped sources.



**Stylianos Ioannis Mimilakis** (S'15) received the Master of Science degree in sound and music computing from the Pompeu Fabra University, Barcelona, Spain and the Bachelor of Engineering in sound and music instruments technologies from the Higher Technological Education Institute of Ionian Islands, Greece. He is currently working toward the Ph.D. degree in signal processing for music source separation, under the MacSeNet project with Fraunhofer Institute for Digital Media Technology, Germany. His research interests include, inverse problems in audio signal processing and synthesis, singing voice separation, and deep learning.

**Derry FitzGerald** received the Ph.D. and M.A. degrees from the Dublin Institute of Technology, Dublin, Ireland and he also received the B. Eng. degree. From 2008 to 2013, he was a Post-Doctoral Researcher with the Department of Electronic Engineering, Cork Institute of Technology, Cork, Ireland, then after that he was a Stokes Lecturer in sound source separation algorithms with the Audio Research Group, DIT. He was a Chemical Engineer with the pharmaceutical industry for some years and also in the field of music and audio, he was a Sound Engineer and has written scores for theatre. He is currently a Research Fellow with the Cork School of Music, Cork Institute of Technology. He has utilized his sound source separation technologies to create the first ever officially released stereo mixes of several songs for the *Beach Boys*, including *Good Vibrations* and *I Get Around*. His research interests include sound source separation and tensor factorizations.

**Bryan Pardo** (M'07) received the M. Mus. degree in Jazz Studies and the Ph.D. degree in computer science, both from the University of Michigan, Ann Arbor, MI, USA, in 2001 and 2005, respectively. While finishing his doctorate, he was with the Department of Music, Madonna University, Livonia, MI, USA and was also a Machine Learning Researcher for General Dynamics, Falls Church, VA, USA. He is currently the Head of the Northwestern University Interactive Audio Lab, Evanston, IL, USA, and an Associate Professor with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA. He has authored or co-authored more than 80 peer-reviewed publications. He has developed speech analysis software for the Department of Speech and Hearing, Ohio State University, Columbus, OH, USA, and a statistical software for SPSS (a software package). When he's not programming, writing or teaching, he performs throughout the United States on Saxophone and Clarinet at venues such as Albion College, the Chicago Cultural Center, the Detroit Concert of Colors, Bloomington Indiana's Lotus Festival, and Tucson's Rialto Theatre.