Estefanía Cano, Derry FitzGerald, Antoine Liutkus,
Mark D. Plumbley, and Fabian-Robert Stöter

Musical Source Separation

An introduction

©ISTOCKPHOTO.COM/TRAFFIC_ANALYZER

Many people listen to recorded music as part of their everyday lives, e.g., from radio or TV programs, compact discs, downloads, or, increasingly, online streaming services. Sometimes we might want to remix the balance within the music, perhaps to make the vocals louder or to suppress an unwanted sound, or we might want to upmix a two-channel stereo recording to a 5.1-channel surround sound system. We might also want to change the spatial location of a musical instrument within the mix. All of these applications are relatively straightforward, provided we have access to separate sound channels (stems) for each musical audio object.

However, if we only have access to the final recording mix, which is usually the case, this is much more challenging. To estimate the original musical sources, which would allow us to remix, suppress, or upmix the sources, we need to perform musical source separation (MSS).

In the general source separation problem, we are given one or more mixture signals that contain different combinations of some original source signals. This is illustrated in Figure 1, where four sources, i.e., vocals, drums, bass, and guitar, are all present in the mixture. The task is to recover one or more of the source signals given the mixtures. In some cases, this is relatively straightforward, e.g., if there are at least as many mixtures as there are sources and if the mixing process is fixed, with no delays, filters, or nonlinear mastering [1].

However, MSS is normally more challenging. Typically, there may be many musical instruments and voices in a two-channel recording, and the sources have often been processed with the addition of filters and reverberation (sometimes nonlinear) in the recording and mixing process. In some cases, the sources may move or the production parameters may change, meaning that the mixture is time varying.

Nevertheless, musical sound sources have particular properties and structures that can help us. For example, musical source signals often have a regular harmonic structure of frequencies at regular intervals and can have frequency contours characteristic of each musical instrument. They may also repeat, in particular, temporal patterns based on the musical structure.

Digital Object Identifier 10.1109/MSP.2018.2874719
Date of publication: 24 December 2018

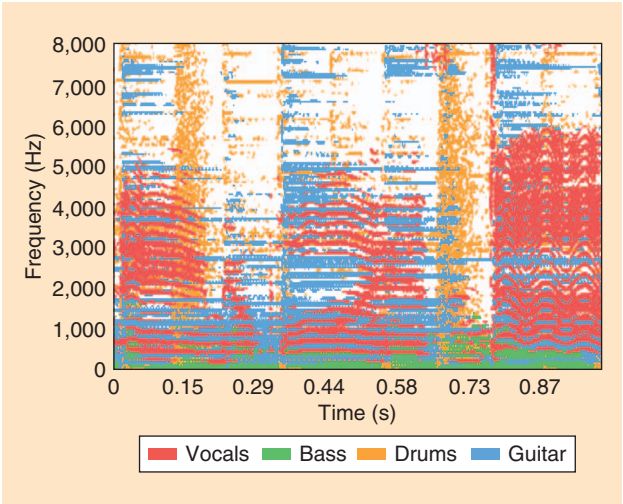

FIGURE 1. A representation of a music mixture in the TF domain. The dominant musical source in each TF bin is displayed with a different color.

In this article, we explore the MSS problem and introduce approaches to tackle it. We begin by presenting characteristics of music signals; we then introduce MSS and, finally, consider a range of MSS models. We also discuss how to evaluate the MSS approaches and discuss limitations and directions for future research. (A dedicated website with complementary information about MSS including sound examples, an extended bibliography, data set information, and accompanying code can be accessed at https://soundseparation.songquito.com/.)

Characteristics of music signals

Music signals have distinct characteristics that clearly differentiate them from other types of audio signals, such as speech or environmental sounds. These unique properties are often exploited when designing MSS methods, and so an understanding of these characteristics is crucial.

All music separation problems start with the definition of the desired musical source to be separated, often referred to as the *target source*. In principle, a musical source refers to a particular musical instrument, such as a saxophone or a guitar, that we wish to extract from the audio mixture. In practice, the definition of musical source is often more relaxed and can refer to a group of musical instruments with similar characteristics that we want to separate. This is the case, e.g., in singing voice separation, where the goal often includes the separation of both the main and background vocals from the mixture. In some cases, the definition of musical source can be even looser, as is the case in harmonic–percussion separation, where the aim is to separate the pitched instruments from the percussive ones.

In a general sense, musical sources are often categorized as predominantly harmonic, predominantly percussive, or as singing voice. Harmonic music sources mainly contain tonal components and, as such, are characterized by the pitch or pitches they produce over time. Harmonic signals exhibit a clear structure composed of a fundamental frequency F0 and a harmonic series. For most instruments, the harmonics appear at multiple integers of the fundamental frequency: for a given F0 at 300 Hz, a harmonic component can be expected close to 600 Hz, the next harmonic component around 900 Hz, and so on. Nonetheless, a certain degree of inharmonicity, i.e., deviation of harmonics from multiple integers of F0, should be expected and accounted for. Harmonic sources exhibit a relatively stable behavior over time and can typically be identified in the spectrogram as horizontal components. This can be observed in Figure 2, where a series of notes played by an acoustic guitar are displayed. Additionally, the trajectories in time of the F0 and the harmonics are usually very similar, a phenomenon referred to as *common fate* [2]. This can be clearly seen in the spectrogram of the vocals in Figure 2, where it can be observed that the vocal harmonics have common trajectories over time. The relative strengths of the harmonics, and the way that the harmonics evolve over time, contribute to the characteristic sound, or timbre, which allows the listener to differentiate one instrument from another. Furthermore, and particularly in Western music, the sources often play in harmony, where the ratios of the F0s of the notes are close to integer ratios. While harmony can result in homogeneous and pleasing sounds, it most often also implies a large degree of overlap in the frequency content of the sources, which makes separation more challenging.

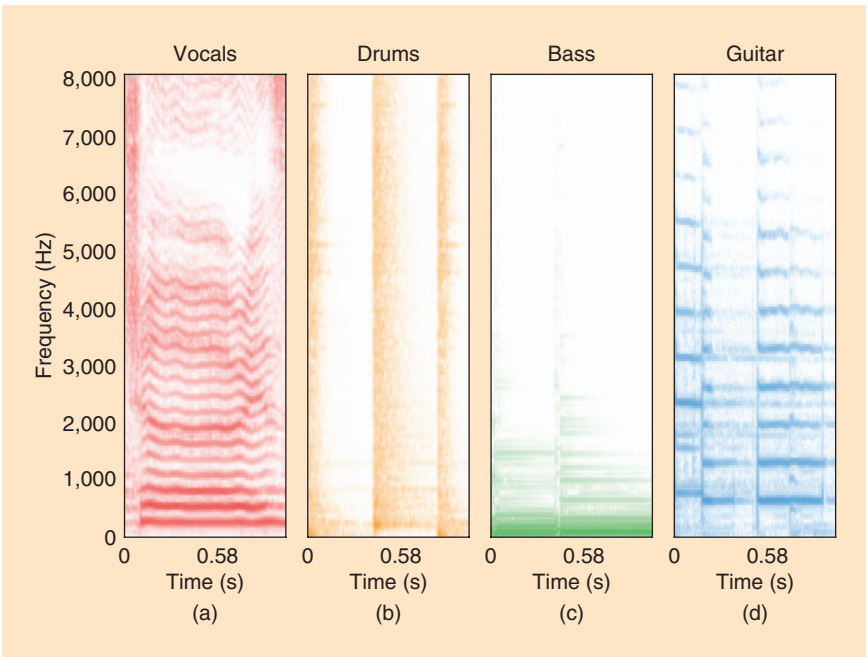

FIGURE 2. A magnitude spectrogram of four example music signals: (a) vocals, (b) drums, (c) bass, and (d) acoustic guitar.

In contrast to harmonic signals, which contain only a selected number of harmonic components, percussive signals contain energy in a wide range of frequencies. Percussive signals exhibit a much flatter spectrum and are highly localized in time with transient-like characteristics. This is apparent in the drums spectrogram in Figure 2, which shows clear vertical structures produced by the drums signal. Percussive instruments play a very important role of conveying rhythmic information in music, giving a sense of speed or tempo in a musical piece.

In reality, most music signals contain both harmonic and percussive elements. For example, a note produced by a piano is considered predominantly harmonic, but it also contains a percussive attack produced by the hammer hitting the strings. Similarly, singing voice is an intricate combination of harmonic (voiced) components, produced by the vibrations of the vocal chords and percussive-like (plosive) components including consonant sounds such as "k" or "p," where no vocal fold vibration occurs. These components are, in turn, filtered by the vocal cavity, with different formant frequencies created by changing the shape of the vocal cavity. As shown in Figure 2, singing voice typically exhibits a higher rate of pitch fluctuation compared with other musical instruments.

A notable property of musical sources is that they are typically sparse in the sense that for the majority of points in time and frequency, the sources have very little energy present. This is commonly exploited in MSS and can be clearly seen for each of the sources in Figure 2. Another characteristic of music signals is the fact that they typically exhibit repeating structures over different time scales, e.g., a repeating percussion pattern over a couple of seconds, to larger structures such as the verse-chorus structures found in pop songs. As explained in the section "Musical Source Models," these repetitions can be leveraged when performing MSS.

The bulk of research on MSS to date has focused on Western pop music as the main genre to be separated. It should be noted that other types of music present their own unique problems for MSS, e.g., many instruments playing in unison in some types of traditional/folk music. These cases are typically not covered by existing MSS techniques.

Once the target source has been defined, the characteristics of the audio mixture should be carefully considered when developing MSS methods. Modern music production techniques offer innumerable possibilities for transforming and shaping an audio mix. Most music signals today are created using audio content from a great diversity of origins, usually combined and mixed using a digital audio workstation (DAW), a software system for transforming and manipulating audio tracks. As depicted in Figure 3 for the trumpet, some musical sources can be recorded in a traditional manner using a microphone or a set of microphones. Very often, hardware devices that shape and color the sound are introduced into the signal chain; these may include, e.g., guitar tube amplifiers or distortion pedals that can impart very particular sound qualities to the signal. In other cases, musical sources are not captured using a microphone. Instead, the digital signal they produce is directly fed into the DAW, or alternatively created within the system itself, using, e.g., a keyboard as an interface as shown in Figure 3. Most frequently, an audio interface is used to facilitate the process of capturing input signals from different origins and delivering them to the DAW. The DAW itself offers many additional possibilities to further enhance and modify the signal.

Once all the audio content is in the DAW, the final step is the creation of the audio mixture. Most commercial music today is in stereo format (two channels). Other multichannel formats such as 5.1 (five main channels plus a low-frequency channel) are available but less common. The number of channels available to perform music separation is a key factor that can be exploited when designing MSS models. Multichannel mixtures allow spatial positioning of the music sources in the sound field. This means that a certain source can be perceived as coming from the left, the center, the right, or somewhere in between. The spatial positioning of the sources is usually achieved with a pan pot that regulates the contribution of each musical source in each of the available channels. This artificial creation of spatial positioning ignores delay as a spatial cue, and so interchannel delay is much less important in MSS than in speech source separation. In contrast, monophonic (single-channel) recordings offer no information about spatial positioning and are often the most challenging separation problem.

A final but fundamental aspect to be considered when designing MSS systems is the fact that music quality, and audio quality in general, is inherently defined and measured

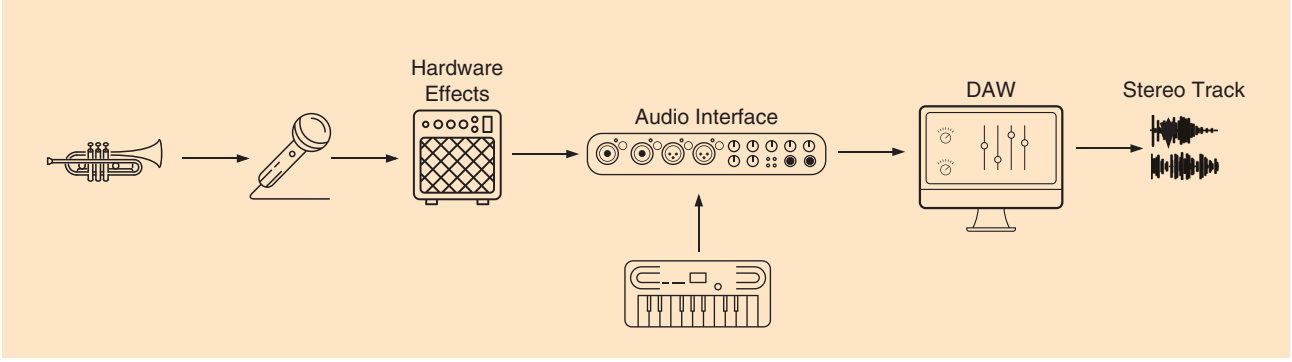

FIGURE 3. The common music recording and mixing setup. In most cases, musical content is combined and mixed into a stereo signal using a DAW.

by human perception. This sets an additional challenge to MSS methods: regardless of the mathematical soundness of the model, all systems are expected to result in perceptually pleasing music signals. Aside from the task of truthfully capturing the target source, we must also minimize the impact on perceptual quality of the distortions introduced in the separation process.

A typical MSS workflow

A high-level overview of the steps involved in most MSS systems is illustrated in Figure 4. First, the input mixture signal is transformed to the time–frequency (TF) domain. The TF representation of the signal is then manipulated to obtain parameters that model the individual sources in the mixture. These are then used to create filters to yield TF estimates of the sources. This is typically done in an iterative manner before the final estimated time-domain signals are recovered via an inverse TF transform. To describe these steps in greater detail, we first introduce the mathematical notation used.

Notation

MSS involves decomposing a time-domain audio mixture signal x into its constituent musical sources y_j. Both x and y_j are vectors of samples in time. Here, the index j denotes the musical source, with $j \in \{1 \cdots J\}$ and J the total number of sources in the mixture. TF representations are denoted in uppercase, with X denoting the complex spectrogram of the mixture x and Y_j denoting the complex spectrogram of the source y_j. Round brackets are used to denote individual elements in a TF representation, with frequency bins denoted with k and time frames denoted with n. The magnitude spectrogram of source j is defined as $S_j = |Y_j|$, whereas $\hat{Y}_j$, $\hat{S}_j$, and y_j denote estimates of the source TF representation, magnitude spectrogram, and source signal, respectively.

Interchannel delay is much less important in MSS than in speech source separation.

TF transformation

Most research in MSS has focused on the use of the short-time Fourier transform (STFT), $X(k, n) \in \mathbb{C}$. The complex STFT has the advantage of being computationally efficient, invertible, and linear: the mixture equals the sum of the sources in the transformed domain, $X = \Sigma_j Y_j$. Other transform alternatives like the constant Q transform have been proposed but have not, as yet, found widespread use in MSS.

Source modeling

Most MSS methods focus solely on analyzing the magnitude spectrogram $M = |X|$ of the mix. The goal at this stage is to estimate either a model of the spectrogram of the target source or a model of the location of the target source in the sound field. As explained in the "MSS Models" section, source and spatial models are the most common approaches used for MSS. Figure 4 shows an example where, starting with the mixture X, estimates of the magnitude spectrograms of the sources $\hat{S}_j$ are obtained.

Filtering

The goal at this stage is to estimate the separated music source signals given the source models. This is typically done using a soft-masking approach, the most common form of which is the generalized Wiener filter (GWF) [3], although other soft-masking approaches have been used [4]. Given X and $\hat{S}_j$, this allows recovery of the separated sources provided their characteristics are well estimated. The STFT of source $j = 1$ can be estimated elementwise using the GWF as $\hat{Y}_1(k, n) = X(k, n)\hat{S}_1(k, n)/\Sigma_j \hat{S}_j(k, n)$. The same procedure is applied for all sources in the mix. Essentially, each TF point in the original mixture is weighted with the ratio of the source magnitude to the sum of the magnitudes of all sources. This can be understood as a multiband equalizer of hundreds of bands, changed dynamically every few milliseconds to attenuate or let pass the required frequencies for the desired source. A special case of the GWF is the process of binary masking, where it is assumed that only one source has energy at a given TF bin, so that mask values are either 0 or 1.

Estimating source parameters from the mixture is not trivial, and it can be difficult to obtain good source parameters to enable successful filtering at the first try. For this reason, it is common to proceed iteratively. First, separation is achieved with the current estimated source models. These models are then updated from the separated signals, and the process is repeated as necessary. This approach is illustrated in Figure 4 and rooted in the expectation–maximization algorithm. Most of the models presented in the "MSS Models" section may be used in this iterative methodology.

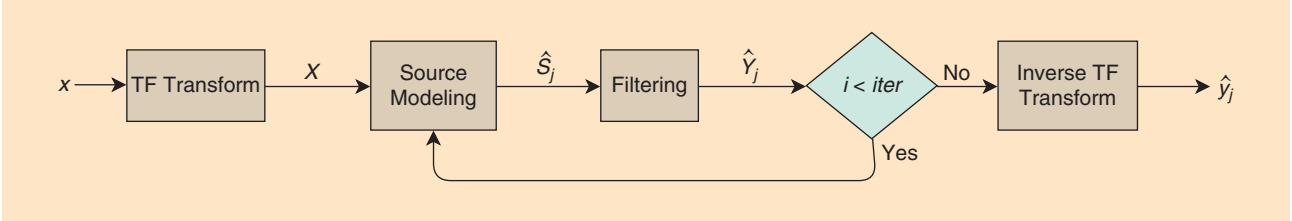

FIGURE 4. The common MSS workflow: source models are obtained from the spectrogram of the audio mix. This is often done in an iterative manner where *iter* represents the total number of iterations and i is the iteration index.

Inverse transform

The final stage in the MSS work flow is to obtain the time-domain source waveforms y_j using the inverse transform of the chosen TF representation.

MSS models

Having described the necessary steps for MSS, we now focus on how the unique characteristics of musical signals are used to perform MSS. While numerous categorizations of MSS algorithms are possible, such as categorization by source type, here we take the approach of dividing the algorithms into two broad categories: algorithms that model the musical sources and those that model the position of the sources in multichannel or stereo audio signals. The key distinction between these two categories is that algorithms in the first category model aspects of the mixture intrinsic to the sources, while those in the second category model aspects intrinsic to the recording/mixing process. Models in the two categories exploit distinct but complementary information that can readily be combined to yield more powerful MSS models.

Musical source position models

In the case of multichannel music signals, the spatial position of the sources has often been exploited to perform music source separation. In this section, we assume that we are dealing with a stereo (two-channel) mixture signal and that the spatial positioning of source j has been achieved using a constant power panning law, defined by a single parameter, the panning angle $\phi_j \in [0, \pi/2]$. For a given source q_j, its stereo representation (or stereo image) is given by $y_{1j} = q_j \cos\theta_j$ and $y_{2j} = q_j \sin\theta_j$, with the subscripts 1 and 2 explicitly denoting the first and second channels, respectively. Figure 5 illustrates the spatial positioning of three sources. The singing voice is positioned in the center and, hence, its stereo image is obtained with an angle of $\pi/4$.

One of the earliest techniques used to exploit spatial position for MSS was independent component analysis (ICA) [1], which estimates an unmixing matrix for the mixture signals based on statistical independence of the sources. However, ICA requires mixtures that contain the same number of channels as musical sources in the mix. This is not typically the case for music signals, where there are usually more sources than channels.

As a result of the shortcomings of ICA, algorithms that worked when there were more sources than channels were developed. Several techniques utilizing spatial position for separation, such as the Degenerate Unmixing Estimation Technique (DUET) [5], Azimuth Discrimination Resynthesis (ADRess) [6], and the PROJection Estimation Technique (PROJET) [7], assume that the TF representations of the sources have very little overlap. This assumption, which holds to a surprising degree for speech, does not hold entirely for music, where the use of harmony and percussion instruments ensures there is overlap. Nonetheless, this assumption has proved to be useful in many circumstances and often results in fast algorithms capable of real-time separation. The degree of overlap between sources can be seen in the spectrogram shown in Figure 1, which only shows the dominant musical source in each TF bin of the mixture.

To illustrate the usefulness of assuming very little overlap between sources, consider the elementwise ratio of the individual mixture channels in the TF domain $R(k, n) = |X_1(k, n)/X_2(k, n)|$, where the subscript numbers indicate the channel number. If there is little TF overlap between the sources, then a single source j will contribute most of the energy at a single point in the TF representations, and so $R(k, n) \approx \cos\phi_j / \sin\phi_j$. Given that $R(k, n)$ only depends on ϕ_j under this assumption, it can therefore be used to estimate a panning angle for each TF point. By plotting an energy-weighted histogram of these angle estimates, a mixture spatial histogram as shown in Figure 5 can be obtained. A peak in this histogram then provides an estimate of the panning angle ϕ_j of a given source. All TF points with an angle close to that of the jth peak are assigned to source j. Then, recovery of an estimate of source j can be done by means of binary masking. DUET, ADRess, and PROJET all estimate histograms of energy versus angle. The main difference between the techniques lies in how these histograms are generated and used and in the type of masking used. Both DUET and ADRess require peak picking from a mixture spatial histogram and use binary masking. PROJET estimates individual source position histograms, the superposition of which results in the mixture spatial position histogram (as shown in Figure 5). Furthermore, PROJET utilizes the GWF for masking. It should also be noted that both DUET and PROJET can also incorporate and deal with interchannel delays, extending their range of applicability.

The aforementioned separation methods directly model sound-engineering techniques such as panning and delay for creating multichannel mixtures. Another line of research models the

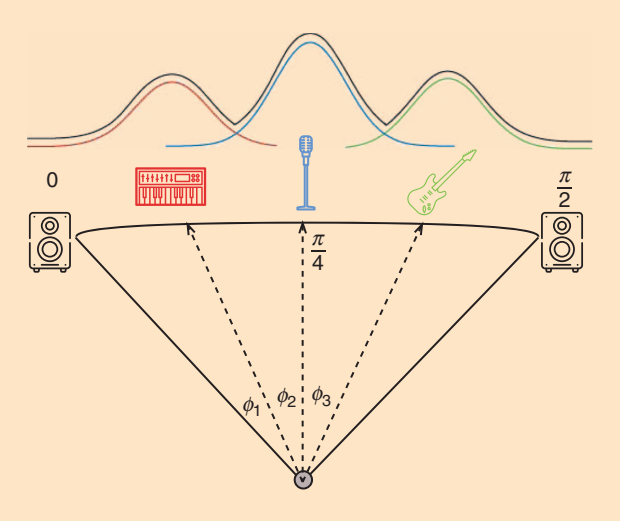

FIGURE 5. An illustration of standard Pan law. The position of the source $j = 1$ (keyboard), source $j = 2$ (voice), and source $j = 3$ (guitar) is defined by the angle ϕ_j, which is always measured with respect to the first channel. Also shown are individual (colored) source position histograms and the mixture spatial histogram (black).

spatial configuration of a source directly through interchannel correlations: at each frequency k, the correlation between the STFT coefficients of the different channels is calculated and encoded in a matrix called the *spatial covariance matrix*. The core idea of such methods is to leverage the correlations between channels to design better filters than those obtained by considering each channel in isolation. This approach is termed *local Gaussian modeling* (*LGM*) [3], [8]. It can give good separation whenever the spatial covariance matrices are well estimated for all sources. It is also able to handle highly reverberated signals, for which no single direction of arrival may be identifiable. This strength of covariance-based methods in dealing with reverberated signals comes at the price of difficult parameter inference. LGM algorithms are often very sensitive to initialization, and the estimated spatial covariance matrices alone are often not sufficient to allow separation. Successful LGM methods need to incorporate musical source models to further guide the separation to obtain acceptable solutions. Musical source models are discussed in the following section.

Musical source models

While spatial positioning can give good results if each source occupies a unique stereo position, it is common for multiple sources to be located in the same stereo position, or for the mixture signal to consist of a single channel. In these cases, model-based approaches that attempt to capture the spectral characteristics of the target source can be used. In the following sections, a range of MSS model-based approaches are described.

Kernel models

Similarly to the idea that the definition of the target sources can be relatively loose in an MSS scenario (see the "Characteristics of Music Signals" section), source models can also incorporate different degrees of specificity when it comes to describing the sources in the mix. Consider the harmonic–percussive separation task: harmonic sources are characterized as horizontal components in the spectrogram, while percussive sources are characterized as time-localized vertical ones. These particularities of the sources can also be understood as harmonic sources exhibiting continuity over time and percussive sources exhibiting continuity over frequency. Music separation models such as kernel additive models (KAMs) particularly exploit local features observable in music spectrograms, e.g., continuity, repetition (at both short and longer timescales such as repeating verses and choruses), and common fate [9]. To estimate a music source at a given TF point, the KAM involves selecting a set of TF bins, which, given the nature of the target source, e.g., percussive, harmonic, or vocals, are likely to be similar in value. This set of TF bins is termed a *proximity kernel*. An example of a vertical proximity kernel used to extract percussive sounds is shown in Figure 6(a). Here, a set of adjacent frequency bins are chosen since percussion instruments tend to exhibit stable ridges across frequency. The vertical kernel is also positioned on a percussive hit seen in the spectrogram of the mixture in Figure 6, where the middle TF bin (k, n) is the one to be estimated. In the case of sustained pitched sounds that tend to have similar values in adjacent time frames, a suitable proximity kernel consists of adjacent time frames across a single frequency [see the horizontal kernel in Figure 6(a)]. KAM approaches assume that interference due to other sources than the target source is sparse, so TF bins with interference are regarded as outliers. To remove these outliers and to obtain an estimate of the target source, the median amplitude of the bins in the proximity kernel is taken as an estimate of the target source at a given TF bin. The median acts as a statistical estimator robust to outliers in energy. Once the

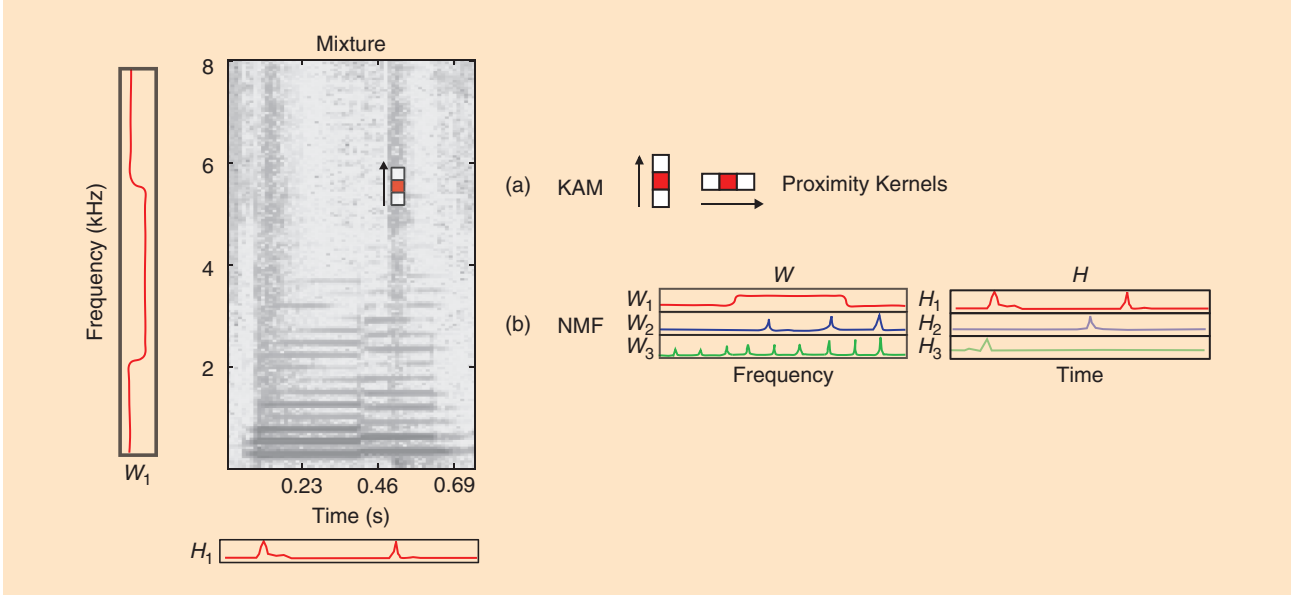

FIGURE 6. An example of different models used in MSS. (a) The proximity kernels used for harmonic–percussive separation within a KAM approach. (b) The spectral templates W and time activations H within an NMF approach.

proximity kernels have been chosen for each of the sources to be separated, separation proceeds iteratively in the manner described in the "A Typical MSS Workflow" section. The KAM is a generalization of previous work on vocal separation and harmonic–percussive separation [4], [10] and has demonstrated considerable utility for these purposes [9].

Spectrogram factorization models

Another group of musical source models is based on spectrogram factorization techniques. The most common of those is nonnegative matrix factorization (NMF). NMF attempts to factorize a given nonnegative matrix into two nonnegative matrices [11]. For the purpose of MSS, NMF can be applied to the nonnegative magnitude spectrogram of the mix M. The goal is to factorize M into a product $M \approx \mathrm{WH}$ of a matrix of basis vectors W, which is a dictionary of spectral templates modeling spectral characteristics of the sources, and a matrix of time activations H. Figure 6(b) shows a series of spectral templates and their corresponding time activations. One of the spectral templates W_1 and its corresponding time activations H_1 are also displayed next to the mixture spectrogram. Peaks in the time activations represent the time instances where a given spectral template has been identified within the mix. Note that the peaks in the time activation H_1 coincide with the percussive hits (vertical structures) in the spectrogram. The factorization task is solved as an optimization problem where the divergence (or reconstruction error) between M and WH is minimized using common divergence measures D such Kullback–Leibler or Itakura–Saito: $\min_{W,H \geq 0} D(M \| WH)$. Many variants of NMF algorithms have been proposed for the purpose of MSS, including methods with temporal continuity constraints [12], multichannel separation models [8], score-informed separation [13], among others [14].

The NMF-based methods described typically assume that the spectrogram for all sources may be well approximated through low-rank assumptions, i.e., as the combination of only a few spectral templates. While this assumption is often sufficient for instrumental sounds, it generally falls short on modeling vocals, which typically exhibit great complexity and require more sophisticated models. In this respect, an NMF variant that has been particularly successful in separating the singing voice uses a source-filter model to represent the target vocals [15]. The idea behind such models is that the voice signal is produced by an excitation that depends on a fundamental frequency (the source), while the excitation is then filtered by the vocal tract or by spectral shapes related to timbre (the filter). A dictionary of source spectral templates and a matrix of filter spectral shapes are used in this model within an expectation–maximization framework.

Some models for the singing voice are based on the observation that spectrograms of vocals are usually sparse, composed of strong and well-defined harmonics, and mostly zero elsewhere (as seen in Figure 2). In this setting, the observed mixture is assumed to be equal to the accompaniment for a large portion of the mixture spectrogram entries. This is the case of robust principal component analysis [16], which does not rely on overconstraining low-rank assumptions on the vocals and, in turn, uses the factorization only for the accompaniment, leaving the vocals unconstrained. Many elaborations on this technique have been proposed. For instance, [17] incorporates voice activity detection in the separation process, allowing the vocals to be inactive in segments of the signal and thus strongly boosting performance.

Sinusoidal models

Another strand of research for MSS models focuses on sinusoidal modeling. This method works under the premise that any music signal can be approximated by a number of sinusoids with time-varying frequencies and amplitudes [18]. Intuitively, sinusoidal modeling offers a clear representation of a music signal, which in most cases is composed of a set of fundamental frequencies and their associated harmonic series. If the pitches present in the target source, as well as the spectral characteristics of the associated harmonics of each pitch, are known or can be estimated, sinusoidal modeling techniques can be very effective for separation of harmonic sources. However, given the complexity of the model and the very detailed knowledge of the target source required to successfully create a realistic representation, the use of sinusoidal modeling techniques for MSS has been limited. Sinusoidal modeling techniques have been proposed for harmonic sound separation [19] and harmonic–percussive separation [20].

Deep neural network models

Historically, MSS research has focused heavily on the use of model-based estimation that enforced desired properties on the source spectrograms. However, if the properties required by the models are not present, separation quality can rapidly degrade. More recently, the use of deep neural networks (DNNs) in MSS has rapidly increased.

In contrast to the approaches described previously, which require explicit models of the source for processing, methods based on DNNs take advantage of optimization techniques to train source models in a supervised manner [21]–[23], i.e., using data sets where both the mix and the isolated sources are available. As depicted in Figure 7, most supervised DNN-based methods take magnitude spectrograms of the audio mix as inputs, optionally also incorporating some further context cues. The targets are set either as the magnitude spectrograms *Sj* of the desired sources (also shown in Figure 7) or as their separating masks (either soft masks or binary masks as described in the section "A Typical MSS Workflow"). Regardless of the inputs and targets used, DNN methods work by training the parameters of nonlinear functions to minimize the reconstruction error of the chosen outputs (spectrograms or masks) based on the inputs (audio mixes).

The models obtained from a DNN depend on two core aspects. First, the type and quantity of the data used for training is of primary importance. To a large extent, representative training data overcome the need for explicitly modeling the underlying spectral characteristics of the musical sources: they

are directly inferred by the network. Second, the DNN topology is of great significance, both for the training capabilities of the network and as a means of incorporating prior knowledge in the system.

The earliest DNN-based approaches for MSS consisted of taking a given frame of the spectrogram, as well as additional context frames as input, and outputting the corresponding frame for each of the targets. These systems mostly consisted of fully connected networks (FCNs). However, given the large size of music spectrograms, the resulting FCNs contained a large number of parameters. This restricted the use of temporal context in such networks to less than 1 s [22]. Therefore, these networks were typically applied on sliding windows of the mixture spectrogram. To overcome this limitation, both recurrent NNs (RNNs) and convolutional NNs were investigated as they offered a principled way to learn dynamic models. RNNs are similar to FCNs, except that they apply their weights recursively over an input sequence and can process sequential data of arbitrary length. This makes such dynamical models ideally suited for the processing of spectrograms of different tracks with different durations, while still modeling long-term temporal dependencies. The most commonly used setting for DNN-based separation today is to fix the number and the nature of the sources to separate beforehand. For example, we may learn a network able to separate drums, vocals, and bass from a mixture. However, having MSS systems that can dynamically detect and separate an arbitrary number of sources is an open challenge, and deep clustering [24] methods represent a possible approach to designing such systems.

A limitation of many DNN models is that they often use mean squared error (MSE) as a cost function. While MSE results in a well-behaved stochastic gradient optimization problem, it also poorly correlates with perceived audio quality. This is the reason why the design of more appropriate cost functions is also an active research topic in MSS [23].

Finally, a crucial limitation of current research for DNN-based MSS is the need for large amounts of training data. The largest MSS multitrack public data set today is MUSDB (https://doi.org/10.5281/zenodo.1117372), which comprises 10 h of data (refer to the accompanying website for more information about available data sets: https://soundseparation.songquito.com/evaluationOf.htm). However, this is still small compared with existing speech corpora comprising hundreds or thousands of hours. The main difficulty in creating realistic multitrack data sets for MSS comes from the fact that individual recordings for each instrument in a mixture are rarely available. Conversely, if individual recordings of instruments are available, the process of creating a realistic music mixture with them is time consuming and very costly (as outlined in the "Characteristics of Music Signals" section). As a result, designing DNN architectures for MSS still requires fundamental knowledge about the sources to be separated, their input and output representations, as well as the use of suitable signal processing techniques and postprocessing operations to further improve the recovered sources.

Evaluation of MSS models

Once an MSS system has been developed, its performance needs to be evaluated. All MSS approaches invariably introduce unwanted artifacts in the separated sources. These artifacts can be caused by mismatches between the models and the sources, musical noise that appears whenever rapid phase or spectral changes are introduced in the estimates (e.g., when using binary masking), reconstruction errors in resynthesis, as well as phase errors in the estimates. Quality evaluation in MSS systems is, however, nontrivial. As musical signals are intended to be heard by human listeners, it is therefore reasonable that evaluation should be based on subjective listening tests such as the Multiple Stimulus with Hidden Reference and Anchors test [25]. However, listening tests are time-consuming and costly, requiring human volunteers to take the tests, and need certain

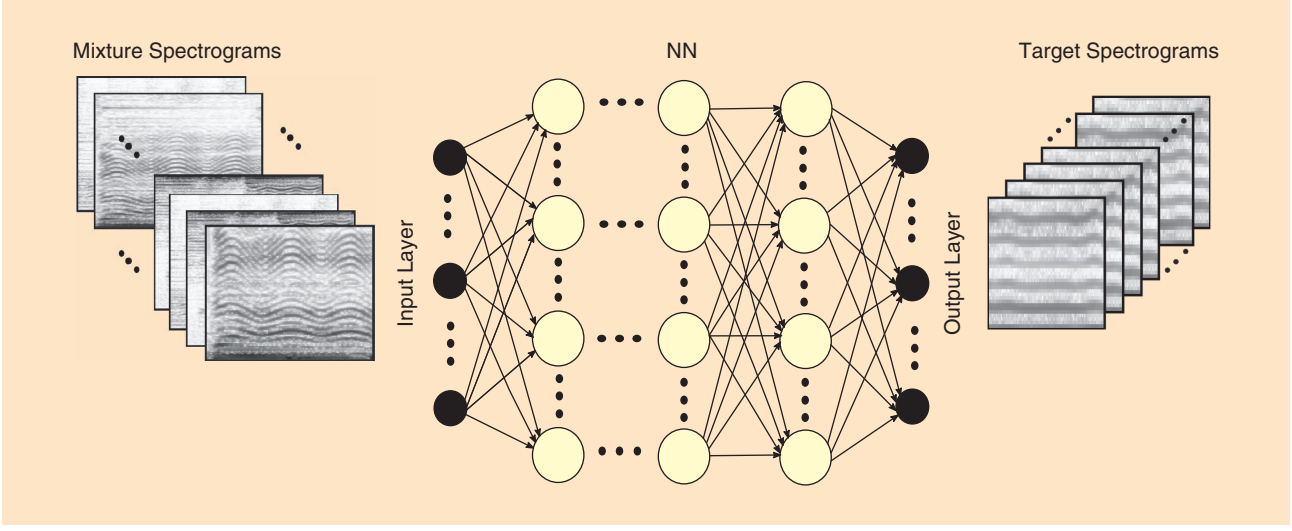

FIGURE 7. The DNN architecture for MSS. The mixture magnitude spectrograms are set as inputs, and source magnitude spectrograms of the desired source S_j are set as the targets.

expertise to be conducted properly. This has made them an infrequent choice for separation quality evaluation.

In an attempt to reduce the efforts of evaluating MSS systems, objective quality metrics have been proposed in the literature. These include the Blind Source Separation Evaluation (BSS Eval) toolbox [26], based on nonperceptual energy ratios, and the Perceptual Evaluation Methods for Audio Source Separation toolkit [27], which aimed to map results obtained via listening tests to create metrics. However, the validity of these metrics has been questioned in recent years, as the values obtained with them do not seem to correlate with perceptual measures obtained via listening tests [28].

DNNs have introduced a paradigm shift in MSS research, with an increasing focus on data-driven models.

Today, given the lack of a unified and perceptually valid strategy for MSS quality evaluation, most algorithm development is still conducted using BSS Eval; however, it is highly recommended to conduct a final listening test to verify the validity of separation results. As a final note, the source separation community runs a regular Signal Separation Evaluation Campaign (SiSEC) [29] including musical audio source separation tasks. SiSEC raises the visibility and importance of evaluation and acts as a focus for discussions on evaluation methodologies.

Future research directions

MSS is a challenging research area with numerous real-world applications. Due to both the nature of the musical sources and the very particular processes used to create music recordings, MSS has many unique features and problems that make it distinct from other types of source separation problems. This is further complicated by the need to achieve separations that sound good in a perceptual sense.

While the quality of MSS has greatly improved in the last decade, several critical challenges remain. First, audible artifacts are still produced by most algorithms. Possible research directions to reduce artifacts include the use of phase retrieval techniques to estimate the phase of the target source, the use of feature representations that better match human perception, allowing models to concentrate on the parts of the sounds that are most relevant for human listeners, and the exploration of MSS systems that model the signal directly in the time domain as waveforms.

Many remaining issues in MSS come from the fact that systems are often not flexible enough to deal with the richness of musical data. For example, it is typically assumed that the actual number of musical sources in a given recording is known. However, this assumption can lead to problems when the number of sources changes over the course of the training procedure. Another issue comes with the separation of sources from the same or similar instrument families, such as the separation of multiple singing voices or violin ensembles.

As previously mentioned, a unified, robust, and perceptually valid MSS quality evaluation procedure does not yet exist. Even while new alternatives for evaluation have been explored in recent years [30], listening tests remain the only reliable quality evaluation method to date. The design of new MSS quality evaluation procedures that are applicable for a wide range of algorithms and musical content will require large research efforts such as large-scale listening experiments, common data sets, and the availability of a wide range of MSS algorithms for use in development.

Additionally, a better understanding of how DNN-based techniques can be exploited for music separation is still needed. In particular, we need better training schemes to avoid overfitting and architectures suitable for music separation. The inclusion of perceptually based optimization schemes and availability of training data are also current challenges in the field.

Recent developments in the area of DNNs have introduced a paradigm shift in MSS research, with an increasing focus on data-driven models. Nonetheless, previous techniques have achieved considerable success in tackling MSS problems. We believe that combining the insights gained from previous approaches with data-driven DNN approaches will allow future researchers to overcome current limitations and challenges in MSS.

Acknowledgments

Mark D. Plumbley was partly supported by grants EP/L027119/2 and EP/N014111/1 from the U.K. Engineering and Physical Sciences Research Council and European Commission H2020 "AudioCommons" research and innovation grant 688382. Antoine Liutkus and Fabian-Robert Stöter were partly supported by the research program "KAMoulox" (ANR-15-CE38-0003-01) funded by the L'Agence Nationale de la Recherche, the French state agency for research.

Authors

Estefanía Cano (cano@idmt.frauhofer.de) received her B.Sc. degree in electronic engineering from the Universidad Potificia Bolivariana, Medellín-Colombia, in 2005, her B.A. degree in music–saxophone performance from the Universidad de Antioquia. Medellín-Colombia, in 2007, her M.Sc. degree in music engineering from the University of Miami, Florida, in 2009, and her Ph.D. degree in media technology from the Ilmenau University of Technology, Germany, in 2014. In 2009, she joined the Semantic Music Technologies group at the Fraunhofer Institute for Digital Media Technology, Germany, as a research scientist. In 2018, she joined the Music Cognition Group at the Agency for Science, Technology, and Research in Singapore. Her research interests include sound source separation, analysis and modeling of musical instrument sounds, and the use of music information retrieval techniques in musicological analysis.

Derry FitzGerald (derry.fitzgerald@cit.ie) received his B.Eng. degree in chemical engineering from the Cork Institute of Technology, Ireland, in 1995 and his M.A. degree in music technology and Ph.D. degree in digital signal processing both from the Dublin Institute of Technology, Ireland, in 2000 and 2004, respectively. He has worked as a research fellow at both Cork and Dublin Institutes of Technology and is currently the chief technology officer at AudioSourceRE, an Irish-based

start-up company developing sound source separation technologies. His research interests are in the areas of sound source separation and tensor factorizations.

Antoine Liutkus (antoine.liutkus@inria.fr) received his state engineering degree from Telecom ParisTech, France, in 2005, his M.Sc. degree in acoustics, computer science, and signal processing applied to music (Acoustique, Traitement du Signal, Informatique, Appliqués à la Musique) from the Université Pierre et Marie Curie (Paris VI), France, in 2005, and his Ph.D. degree in electrical engineering from Telecom ParisTech in 2012. He worked as a research engineer on source separation at Audionamix, Paris, France, from 2007 to 2010. He is currently a researcher in the speech processing team at Inria Nancy Grand Est located in Villers-lès-Nancy, France. His research interests include audio source separation and machine learning.

Mark D. Plumbley (m.plumbley@surrey.ac.uk) received his Ph.D. degree in neural networks from the Engineering Department at Cambridge University, United Kingdom, in 1991 and became a lecturer at King's College London, United Kingdom. He moved to Queen Mary University, London, United Kingdom, in 2002, later becoming a professor of machine learning and signal processing and the director of the Centre for Digital Music. In 2015, he joined the University of Surrey, United Kingdom, as a professor of signal processing in the Centre for Vision, Speech, and Signal Processing. His research interests include the analysis and processing of audio and music and using a wide range of signal processing techniques, including independent component analysis and sparse representations.

Fabian-Robert Stöter (fabian-robert.stoter@inria.fr) received his diploma degree in electrical engineering from the Leibniz University of Hanover, Germany, in 2012 and worked toward his Ph.D. degree in audio signal processing in the research group of B. Edler at the International Audio Laboratories Erlangen, Germany. He is currently a researcher at Inria/Laboratory of Computer Science, Robotics, and Microelectronics in Montpellier, France. His research interests include supervised and unsupervised methods for audio source separation and signal analysis of highly overlapped sources.

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[2] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press, 1990.

[3] 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 Language Processing*, vol. 18, no. 7, pp. 1830–1840, Sept. 2010.

[4] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (REPET): A simple method for music/voice separation," *IEEE Trans. Audio Speech Language Processing*, vol. 21, no. 1, pp. 73–84, Jan. 2013.

[5] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation, Signals and Communication Technology*, S. Makino, H. Sawada, T. W. Lee, Eds. Dordrecht, Netherlands: Springer, 2007, pp. 217–241.

[6] D. Barry, B. Lawlor, and E. Coyle, "Real-time sound source separation using azimuth discrimination and resynthesis," in *Proc. 117th Audio Engineering Society (AES) Conv.*, San Francisco, CA, 2004, pp. 1–7.

[7] D. FitzGerald, A. Liutkus, and R. Badeau, "Projection-based demixing of spatial audio," *IEEE/ACM Trans. Audio Speech, Language Processing*, vol. 24, no. 9, pp. 1560–1572, Sept. 2016.

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

[9] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Trans. Signal Processing*, vol. 62, no. 16, pp. 4298–4310, Aug. 2014.

[10] D. FitzGerald, "Harmonic/percusssive separation using median filtering," in *Proc. 13th Int. Conf. Digital Audio Effects (DAFx)*, Graz, Austria, 2010, pp. 1–4.

[11] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances Neural Inform. Processing Syst.*, vol. 13, pp. 556–562, Apr. 2001.

[12] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio Speech Language Processing*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[13] S. Ewert, B. Pardo, M. Müller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Mag.*, vol. 31, no. 3, pp. 116–124, May 2014.

[14] P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Mag.*, vol. 31, no. 3, pp. 66–75, May 2014.

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

[16] 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. Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 57–60.

[17] T.-S. Chan, T.-C. Yeh, Z.-C. Fan, H.-W. Chen, L. Su, Y.-H. Yang, and R. Jang, "Vocal activity informed singing voice separation with the ikala dataset," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 718–722.

[18] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Studies on New Music Research*, M. Leman and P. Berg, Eds. Swets & Zeitlinger, 1997, pp. 91–122.

[19] T. Virtanen and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modeling," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Budapest, Hungary, 2000, pp. II765–II768.

[20] E. Cano, M. Plumbley, and C. Dittmar, "Phase-based harmonic/percussive separation," in *Proc. 15th Annu. Conf. Int. Speech Communication Association Interspeech*, Singapore, 2014, pp. 1628–1632.

[21] 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 Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[22] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 2135–2139.

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

[24] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 31–35.

[25] International Telecommunication Union. (2015, Oct.). Recommendation BS.1534-3: Method for the subjective assessment of intermediate quality levels of coding systems. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1534-3-201510-I/en

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

[27] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio Speech Language Processing*, vol. 19, no. 7, pp. 2046–2057, Sept. 2011.

[28] E. Cano, D. FitzGerald, and K. Brandenburg, "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Budapest, Hungary, 2016, pp. 1758–1762.

[29] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Proc. Int. Conf. Latent Variable Analysis and Signal Separation*, 2017, pp. 323–332.

[30] D. Ward, H. Wierstorf, R.D. Mason, E. M. Grais, and M.D. Plumbley, "BSS Eval or PEASS? Predicting the perception of singing-voice separation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, 2018.

SP