Remote Sensing in Ecology and Conservation
Open Access

ZSL
LET'S WORK FOR WILDLIFE

REVIEW

Source separation in ecoacoustics: A roadmap towards versatile soundscape information retrieval

Tzu-Hao Lin[1] & Yu Tsao[2]

[1]Research Institute for Global Change, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), 2–15 Natsushima-cho, Yokosuka, Kanagawa, 237-0061, Japan
[2]Research Center for Information Technology Innovation, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei, 115, Taiwan

Keywords
acoustic habitat, biodiversity, ecosystem dynamics, machine learning, passive acoustics, signal processing

Correspondence
Tzu-Hao Lin, Research Institute for Global Change, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), 2-15 Natsushima, Yokosuka City, Kanagawa, 237-0061, Japan. Tel: +81 46 867 9691; Fax: +81 46 867 9025;
E-mail: schonkopf@gmail.com

Editor: Nathalie Pettorelli

Received: 29 June 2019; Revised: 24 November 2019; Accepted: 26 November 2019

doi: 10.1002/rse2.141

Abstract

A comprehensive assessment of ecosystem dynamics requires the monitoring of biological, physical and social changes. Changes that cannot be observed visually may be trackable acoustically through soundscape analysis. Soundscapes vary greatly depending on geophysical events, biodiversity and human activities. However, retrieving source-specific information from geophony, biophony and anthropophony remains a challenging task, due to interference by simultaneous sound sources. Audio source separation is a technique that aims to recover individual sound sources when only mixtures are accessible. Here, we review techniques of monoaural audio source separation with the fundamental theories and assumptions behind them. Depending on the availability of prior information about the source signals, the task can be approached as a blind source separation or a model-based source separation. Most blind source separation techniques depend on assumptions about the behaviour of the source signals, and their performance may deteriorate when the assumptions fail. Model-based techniques generally do not require specific assumptions, and the models are directly learned from labelled data. With the recent advances of deep learning, the model-based techniques can yield state-of-the-art separation performance, accordingly facilitate content-based audio information retrieval. Source separation techniques have been adopted in several ecoacoustic applications to evaluate the contributions from biodiversity and anthropogenic disturbance to soundscape dynamics. They can also be employed as nonlinear filters to improve the recognition of bioacoustic signals. To effectively retrieve ecological information from soundscapes, source separation is a crucial tool. We believe that the future integrations of ecological hypotheses and deep learning can realize a high-performance source separation for ecoacoustics, and accordingly improve soundscape-based ecosystem monitoring. Therefore, we outline a roadmap for applying source separation to assist in soundscape information retrieval and hope to promote cross-disciplinary collaboration.

Introduction

Remote sensing techniques are widely employed to identify biophysical characteristics of habitats, as well as to track natural and human-caused impacts on biodiversity (Kerr and Ostrovsky 2003). Visual information alone is unable to address all aspects of ecosystem dynamics. Ecoacoustics is an emerging field that investigates the relationships among sounds of living organisms (biophony), their habitats (geophony) and human development (anthropophony) (Dumyahn and Pijanowski 2011; Pijanowski et al. 2011; Sueur and Farina 2015; Krause and Farina 2016). By listening to biotic sounds, the composition of soniferous animals can be characterized and applied to assess wildlife biodiversity (André et al. 2011; Saito et al. 2015; Lin et al. 2017a; Ross et al. 2018). On the other hand, abiotic sounds can be used to evaluate the ecosystem dynamics associated with geophysical activities and human interference (Guan et al. 2015; Coquereau et al. 2017; Menze et al. 2017). Thus, a soundscape might contain valuable ecological information.

Effectively retrieving information from a soundscape can be very challenging. Sound pressure levels and power spectral densities are commonly used to search sound sources of interest (Guan et al. 2015; Monczak et al. 2019). Regardless of which band of frequencies is chosen to be representative of a particular source, other sound sources will inevitably interfere. To summarize the contributions of biological activities to acoustic variation, many ecoacoustic indices, such as Acoustic Diversity Index, Acoustic Complexity Index, Normalized Difference Soundscape Index, have been developed in terrestrial sound analysis projects (Sueur et al. 2008, 2014; Pieretti et al. 2011; Villanueva-Rivera et al. 2011; Kasten et al. 2012). Many of these indices measure spectral-temporal heterogeneity from a combination of frequencies, and the results are employed to assess biodiversity. However, some indices may be sensitive to changes in the calling rate of a single species, weather conditions or correlated with anthropogenic activities (Fairbrass et al. 2017; Bohnenstiehl et al. 2018). Thus, the utility of these indices as ecological indicators may be limited without removing unwanted sounds.

The difficulty of retrieving source-specific information from geophony, biophony and anthropophony is primarily due to interference by simultaneous sound sources. Sounds rarely occur in isolation, and spectral masking or signal distortion may result in biased measurements, which in turn affect ecological interpretations (Lin et al. 2018). Extensive works have been carried out in recognition of biological sounds by extracting a set of audio features (Lin et al. 2013; Xie et al. 2015; Stowell 2018). Some features are reliable for investigating activity patterns of one or a few species, but there is no guarantee that they can be applied for non-target species or abiotic sounds.

In recent years, speech and musical source separation (SS) have achieved significant advances, but those techniques have not widely applied in ecoacoustics. We noticed that efforts had been dedicated to applying microphone and hydrophone arrays in passive acoustic monitoring of wildlife (Gillespie et al. 2009; Suzuki et al. 2017). With a multi-sensor system to spatially and synchronously record sounds, many techniques, such as beam forming and source localization, are available for the discrimination of species, behaviour and individual identity (Blumstein et al. 2011). However, multi-sensor systems generally demand higher hardware requirements, and their applicability is often restricted to situations where the primary target and interference sources are spatially separated (Wang and Chen 2017). To our knowledge, most soundscape research projects conducted in underwater environments only collect single-channel audio at one recording site, and those techniques for multi-channel audio SS are not usable. Therefore, techniques of monoaural audio SS are essential in the information retrieval of soundscapes.

In this study, we describe a roadmap for applying SS in ecoacoustics. First, we review modern techniques of monoaural audio SS, including blind source separation (BSS) and model-based SS. Then, we investigate how SS can solve problems encountered in ecoacoustics due to interference by simultaneous sound sources. Two functional scales are discussed: the evaluation of soundscape dynamics, and the automated recognition of biological sounds. Finally, we outline items that need to be considered during soundscape separation and future directions for the development of separation tools to facilitate soundscape information retrieval.

Monoaural audio source separation

Monoaural audio SS refers to the technique of extracting one or more source signals of interest from a single-channel mixture of signals. Unlike audio classification, which aims to retrieve categorical labels from source signals (Fu et al. 2011), SS aims to model the temporal variations of source signals in a mixture x_i, with:

$$x_i = \sum_{j=1}^{N} a_{ij} s_j$$

where x_i represents the totality of received signals at time i, S_j represents the signal of source j, and a_{ij} represents the amplitude factor at time i of the signal from source j. Monoaural audio SS is an underdetermined system because a single channel is used to record multiple sources. If the signals of the sources (s) are unknown, then the problem is called BSS, which performs SS in an unsupervised learning manner. If the source signals are (partially) known, then the amplitude factors (α) can be estimated using source-specific models.

Blind source separation

BSS is a technique of decomposing a received mixture of signals without pre-existing labels. Although it is possible to perform monoaural BSS in the time (waveform) domain, more methods operate BSS in the spectral-temporal domain. Namely, the audio waveforms are transformed into time–frequency representations via short-time Fourier transform. The time–frequency representations are subsequently decomposed into a set of components or basis functions, which are used to project the received mixture onto a multi-dimensional feature space. Often, the number of components is chosen as N, according to the number of sources believed to be active. Thus, the mixture can be separated by finding disjoint basis sets for different sources (Fig. 1A). This concept has been applied in methods based on principal component analysis (PCA), independent component analysis (ICA) and non-negative matrix factorization (NMF). Although pre-existing labels are not required, appropriate assumptions regarding the number of sound sources of interest are required (Fig. 1A).

PCA represents the most conventional technique for noise/signal separation. It uses an orthogonal projection to convert a set of correlated features into a set of linearly uncorrelated components. The objective is to identify a sequence of components that accounts for as much data variability as possible. By assuming stationary noise, the components which explain the largest variance are retained to enhance the signals, and noisy components corresponding to the smallest variance are discarded (Vaseghi 2008). Huang et al. (2012) modelled the problem of music/voice separation based on the idea that repetition is a core principle in music. The music background was assumed to display strong repetitiveness, while the singing voice was assumed to exhibit more substantial variation, including moments of relative silence. Their results showed that Robust PCA, a modified PCA which deals with the problem of grossly corrupted observations, could separate the music background and the singing voice by decomposing a mixture spectrogram into a structural low-rank matrix and a sparse matrix.

The assumption of stationary noise may not be reliable in many applications. Unlike PCA, ICA is a method for separating a mixture into a set of statistically independent components by maximizing the non-Gaussianity as the objective function (Comon 1994). ICA-based BSS generally takes multi-channel recordings. Despite that, Virtanen (2006) propose to replace multi-channel inputs as frequency bands derived from fast Fourier transform, or components obtained from PCA to separate signals of different musical instruments from single-channel audio.

NMF is another widely employed machine-learning model in monoaural BSS. With NMF, a non-negative matrix can be decomposed into a set of basis functions

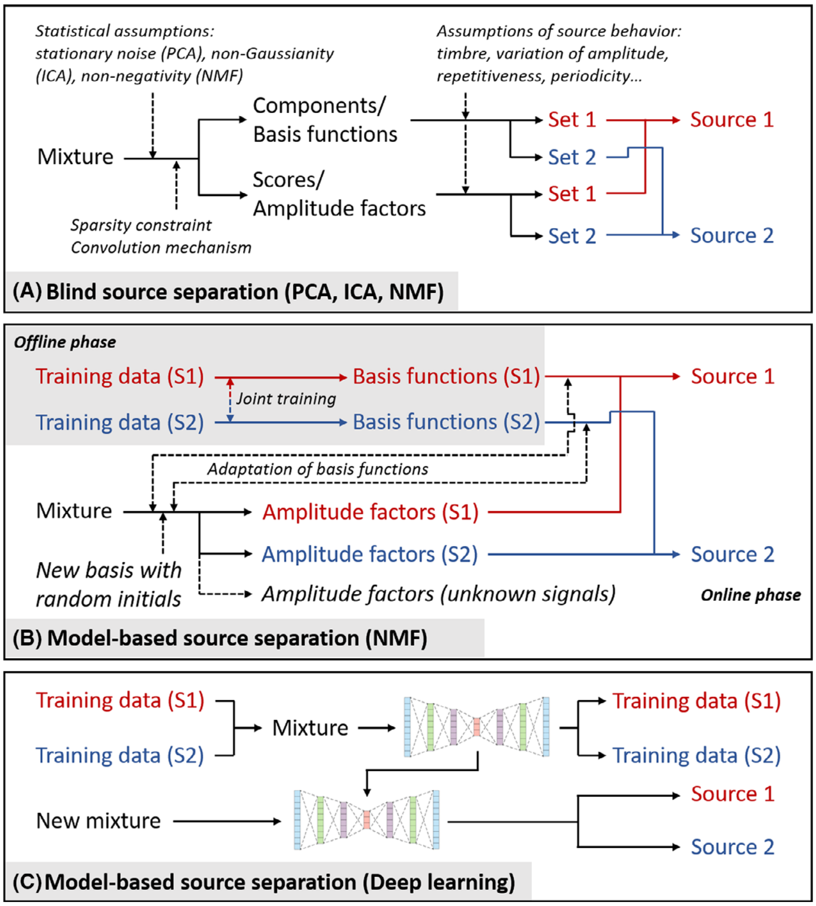
Statistical assumptions:
stationary noise (PCA), non-Gaussianity (ICA), non-negativity (NMF)
Assumptions of source behavior:
timbre, variation of amplitude, repetitiveness, periodicity...
Mixture
Components/ Basis functions
Scores/ Amplitude factors
Sparsity constraint
Convolution mechanism
Set 1
Set 2
Set 1
Set 2
Source 1
Source 2
(A) Blind source separation (PCA, ICA, NMF)
Offline phase
Training data (S1)
Training data (S2)
Joint training
Basis functions (S1)
Basis functions (S2)
Source 1
Adaptation of basis functions
Mixture
Amplitude factors (S1)
Amplitude factors (S2)
Source 2
New basis with random initials
Amplitude factors (unknown signals)
Online phase
(B) Model-based source separation (NMF)
Training data (S1)
Training data (S2)
Mixture
Training data (S1)
Training data (S2)
New mixture
Source 1
Source 2
(C) Model-based source separation (Deep learning)

Figure 1. Comparison of (A) blind source separation, (B) model-based source separation and (C) deep learning-based source separation. Dashed arrows represent the items that need to be considered in monoaural audio source separation.

and associated encoding values by minimizing the reconstruction error. Unlike ICA, NMF does not assume signals to be statistically independent. Instead, NMF requires only that the input be non-negative (Lee and Seung 1999). The non-negative constraint reflects the additive nature of sounds, thus ensuring that the spectral features and temporal strength of source signals are learned in basis functions and encoding values respectively (Smaragdis et al. 2014). The modelling of NMF may incorporate a sparsity constraint, a commonly used attribute for constraining the probability distribution of zero-valued elements, to improve the likelihood of obtaining independent basis functions (Hoyer 2004). On the other hand, a convolutive NMF, which learns temporally modulated basis functions, may be beneficial for learning source signals with time-varying spectral features (O'Grady and Pearlmutter 2006).

In monoaural BSS, prior information about the number of sources is a critical attribute for learning source-specific basis functions. When such prior information is inaccurate, unsupervised learning of basis functions and associated amplitude factors may not perform well. For example, ICA attempts to maximize the independence of the basis functions, but it does not guarantee complete independence when a complicated mixture is analysed. If prior information about the number of sources is not available or ambiguous, ICA or NMF can be employed to learn an overcomplete dictionary, and then to subsequently use clustering to derive specific sets of basis functions and source-specific independent subspaces (Molla et al. 2008).

Successful applications of clustering-based BSS requires an appropriate assumption about the behaviour of source signals, such as timbre and temporal characteristics. For example, Kameoka et al. (2018) assumed that different sources display their specific timbre patterns, and successfully use Mel-frequency cepstral coefficients to separate basis clusters obtained with NMF for various musical instruments. Innami and Kasai (2012) separated background and event sounds by using time-variant features obtained with NMF. Lin et al. (2017b) assumed that biotic and abiotic sounds display source-specific diurnal patterns, and integrated this concept in the periodicity-coded NMF. Their results showed that a second NMF layer could act as a clustering tool to identify basis clusters of biotic and abiotic sounds, thus enhance biological choruses from terrestrial and marine soundscapes. These examples suggest that the integration of domain knowledge or hypotheses in the clustering of source-specific basis functions is a practical approach for monoaural BSS.

Clustering-based BSS has also been tested to separate source signals with similar frequency ranges but distinct spectral modulation patterns. Deep learning frameworks, such as deep neural networks or deep NMF, can produce an embedding for each time–frequency element in a spectrogram (Hinton and Salakhutdinov 2006). By performing clustering on the embeddings, the voices of multiple speakers could be separated (Hershey et al. 2016; Isik et al. 2016). Deep clustering is a relatively new technique, but it shows promise as a method for performing advanced BSS without assumptions about the behaviour of source signals.

Model-based source separation

In general, audio SS can achieve better performance in a supervised learning manner, where pre-existing labels are available for source signals of interest. With source-specific training data, prior information on the source signals (*s* in Eq. S1) can be obtained and subsequently assist in the estimation of amplitude factors (*a* in Eq. S1) from a recorded mixture.

Both ICA and NMF-based SS can be performed in a supervised learning manner (Jang and Lee 2003; Fan et al. 2014; Lin et al. 2017b). Using NMF-based SS as a representative method, there are four integral steps. The first step applies NMF to learn time-domain or spectral-domain basis functions from a training set. The second step is to repeat the learning procedure for different sound sources and concatenate the results, thus producing a dictionary containing basis functions corresponding to each of the source types analysed. The third step iteratively estimates amplitude factors in a received mixture by fixing the parameters of the concatenated basis functions. Finally, the individual sources are recovered using the source-specific basis functions and the corresponding amplitude factors. The first two steps are usually referred to as the offline phase, while the latter two are called the online phase (Fig. 1B).

In model-based SS, the modelling of basis functions according to source-specific training data in the offline phase represents the most critical step. However, source-specific basis functions may contain unwanted components if a training set contains noise. Discriminative basis functions can be obtained by jointly training on a corpus of sound sources (Wang and Sha 2014; Weninger et al. 2014). In addition to discriminative ability, another issue that may affect the separation performance is the precision of annotating labels in the training data. Due to the expensive and time-consuming process to prepare annotated labels, it is common to prepare only coarse labels determining the presence or absence of target signals. One way to learn from such weakly labelled data is to introduce a binary mask on the activation matrix during joint training so that the model can learn to construct discriminative basis functions for the labels of presence and absence (Sobieraj et al. 2017).

In addition to elevating the discriminative ability in the offline phase, we can consider adaptive approaches in the online phase. Conventionally, amplitude factors are learned by fixing the basis functions. Adaptive methods allow for the update of basis functions according to testing data obtained during the online phase (Wu and Lerch 2015). Thus, the separation can be improved when signals of interest slightly vary among recordings.

Interference due to unknown sources is another common issue in the online phase. A semi-supervised learning approach can solve this issue by adding new basis functions with random initialization to the pool of source-specific basis functions (Bryan and Mysore 2013). Components of unknown sources can be learned through the iterative updates of newly added basis functions if their spectral features are distinct from the trained sources. To ensure the minimal spectral similarity between the basis functions of the known sources and the unknown sources, applying a sparsity constraint in the time domain is a practical approach (Smaragdis et al. 2007).

More recently, the advancement of deep learning has achieved considerable improvements in speech signal processing (Wang and Chen 2017). Deep learning is a technique that maps input data to corresponding output by minimizing an error function through multiple layers of processing units (LeCun et al. 2015). For speech enhancement, a noisy spectrogram is used as the input, and a clean spectrogram is used as the output to train a deep learning model (Lu et al. 2013; Xu et al. 2015). Figure 1C shows the framework to separate two sources by training a model to map a spectrogram of the mixture to clean spectrograms of the two sources (Du et al. 2014; Tu et al. 2014). Furthermore, convolutional neural networks (CNN) has allowed end-to-end mapping without first transforming the waveform to time–frequency representations (Fu et al. 2017, 2018; Stoller et al. 2018a). Another alternative approach is to derive a masking function, such as a binary spectrogram that denotes whether the target signal dominates a time–frequency unit (Wang and Chen 2017). By learning from such time–frequency relationships, deep learning models, including fully connected networks (Wang and Wang 2013), recurrent neural network (Erdogan et al. 2015; Huang et al. 2015), and bidirectional long short-term memory (Chen et al. 2015; Uhlich et al. 2015), and CNN (Hui et al. 2015; Nugraha et al. 2016), are the current state of the art in the separation of speech and music (Stöter et al. 2018).

Although deep learning in a supervised learning manner has made significant achievement, the primary disadvantage of most deep learning-based SS is the need to prepare pairs of mixed and pure source sounds. The most common approach to solve this issue is to synthesize a large number of mixtures based on a database of pure source sounds. Solutions to reduce the dependence on pure source sounds have been proposed by using weakly labelled data to guide the representation learning to induce structure (Stowell and Turner 2015; Ewert and Sandler 2017). Variational autoencoder (VAE) is a deep learning framework for building probabilistic generative models by training an encoder together with a decoder which reconstructs the input data (Kingma and Welling 2014). By forcing the entire encoder–decoder model to learn to reconstruct the input, the encoder can generate embeddings that are associated with the structure of the input data. A non-negative form of VAE can learn class information about the sources without supervised training on pure source sounds (Karamatlı et al. 2018). This idea has not been widely tested, so the performance in other environments remains unclear. Despite that, developing SS models with reduced dependence on pure source sounds is an interesting and challenging problem, which is worth further exploration.

Applications of source separation in soundscape-based ecosystem monitoring

The capability to deliver digital information that could identify various geophysical, biological and anthropogenic events from autonomous recorders represents the primary driver for the recent increase in ecoacoustic applications. To facilitate decision making in conservation management, analyses of soundscapes generally focus on the evaluation of soundscape dynamics, and the automatic recognition of biological sounds.

Soundscape dynamics can generate information relevant to the quality of the acoustic habitat and the community of soniferous animals (Farina 2018). Previous studies have employed PCA to statistically distinguish between acoustic variation related mainly to biological activities or anthropogenic disturbances across temporal and geographical gradients (Kuehne et al. 2013; Guan et al. 2015). On the other hand, Eldridge et al. (2016) introduced BSS based on the probabilistic latent component analysis (closely related to NMF) in the analysis of acoustic diversity contributed by animal vocalizations. This approach has relevance to the analysis of acoustic diversity based on the clustering of sound types (Phillips et al. 2018; Ulloa et al. 2018). Lin et al. (2017a) applied the periodicity-coded NMF to separate biotic and abiotic sounds from long-term spectrograms, and subsequently employed clustering to model seasonal changes in bioacoustic diversity in three forests with different altitudes. Although only relatively few ecoacoustics studies had applied monoaural audio SS, the preliminary progress suggests a great potential to assist in the evaluation of ecosystem dynamics.

In the evaluation of soundscape dynamics, a library of pure source sounds to guide the training of source-specific models is generally lacking. Even though weakly labelled data may be obtainable, the ambiguity of the labelling may be high due to source interferences. Moreover, new classes of sound outside the training set are often encountered in long-duration recordings or when a project focuses on a new study area. Therefore, the application of model-based SS may be challenging. The integration of ecological assumptions, such as diurnal or seasonal periodicities, in BSS techniques, serves as a promising solution. For Figure 2 as an example, the periodicity-coded NMF (Lin et al. 2017a) can be used to decompose a spectrogram of long-duration recordings into biological choruses with a strong diurnal pattern (source 1) and other sound sources without a prominent periodical pattern (source 2). Although not all biotic sounds exhibit the same temporal pattern, SS can still help us identify the sound sources which meet specific ecological assumptions.

Another application of SS is to facilitate the extraction of audio features, detection and classification of biological sounds. Bioacoustic studies have applied statistics-based noise reduction, such as spectral subtraction and Wiener filter, to enhance transient vocalizations (Bardeli et al. 2010; Lin et al. 2013; Xie et al. 2015; Lostanlen et al. 2019). However, statistics-based noise reduction does not handle non-stationary noise well. SS models can act as nonlinear filters, thus improving the precision of acoustic measurements. Using Figure 3 as an example, the presence or absence of different sound sources were manually annotated. By learning from such weakly labelled data, a deep NMF model (Le Roux et al. 2015) can accurately recognize the spectral characteristics for bird and insect calls recorded in a forest. Even when pre-existing labels are not available, studies had demonstrated the feasibility of using BSS to detect cetacean vocalizations, fish choruses and shipping noise from estuarine and deep-sea soundscapes (Lin and Tsao 2018; Lin et al. 2019).

There are other potential ecological applications of SS not addressed here. For example, although new and useful acoustic features have been introduced in bioacoustics and ecoacoustics (Bellisario and Pijanowski 2019; Bellisario et al. 2019a), can we learn biologically meaningful features from the data itself? This is entirely in line with the active research field of unsupervised feature learning in informatics (Längkvist et al. 2014). Another example, given

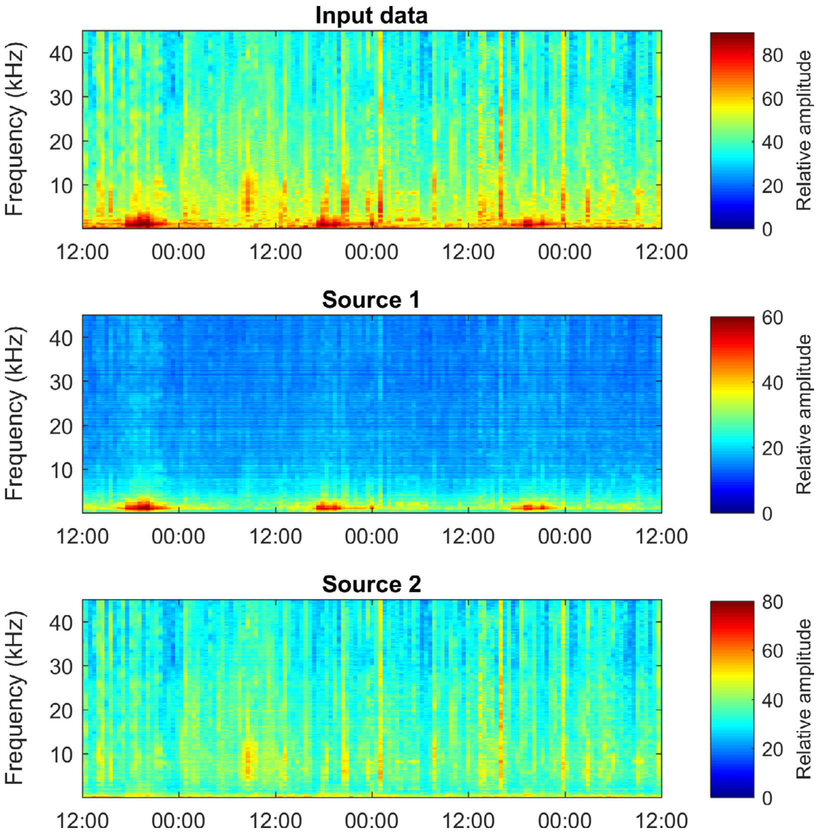

Figure 2. An example of applying monoaural BSS in the analysis of long duration underwater recordings. The input data are a long-term spectral average of underwater recordings collected in Chunggang river estuary, western Taiwan. We assumed that different sound sources display source-specific diurnal patterns and apply periodicity-coded NMF for BSS. The result shows that fish choruses (source 1) and other sound sources (source 2) can be effectively separated.

the prior information on the composition of soniferous animals, how can we obtain a model to predict spatiotemporal changes in the community (Bellisario et al. 2019b)? Besides, SS may be employed to test ecoacoustics hypotheses, including the Acoustic Niche Hypothesis (Krause 1993), Acoustic Adaptation Hypothesis (Ey and Fischer 2009) and Acoustic Habitat Hypothesis (Mullet et al. 2017).

A roadmap to effective retrieval of soundscape information

Many audio applications, including speech, music and soundscape, rely on the ability to search audio content (Wold et al. 1996; Bellisario and Pijanowski 2019; Cano et al. 2019). A soundscape contains multiple dimensions of ecological information, and SS is expected to play a crucial part in assisting information retrieval. Particularly in areas where geophony and anthropophony significantly interfere with soundscapes, SS will be necessary for improving the acoustic assessment of biodiversity.

We reviewed the available techniques of BSS and model-based SS for monoaural audio SS, along with their assumptions and limitations. It is necessary to evaluate the availability of prior information on the sound sources of interest. If labelled data are available, it is possible to train source-specific models. Although model-based SS is powerful, it requires the identification of all sound sources of interest in advance and preparation of sufficient training data for each of them. In many ecosystems, such as tropical forests and deep sea, many biotic and abiotic sounds are yet to be identified. Therefore, a semi-supervised approach may be useful for unknown sound sources. Joint training or adaptation of basis functions can also be carried out to improve the learning of discriminative features.

When labelled data are limited or unavailable, we recommend applying weakly supervised learning or BSS to identify different acoustic events and construct source-specific models, as we demonstrated in Figures 2 and 3. Operators can manually revise the model parameters (or

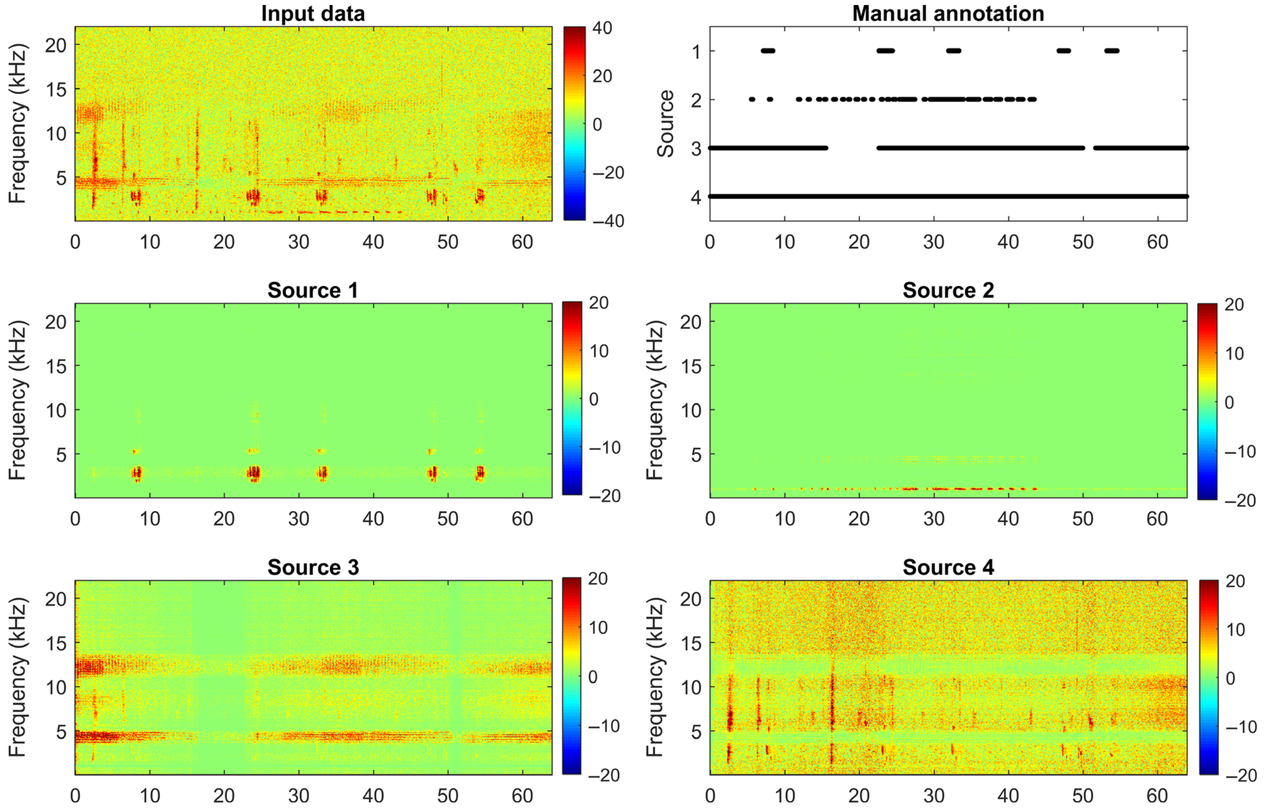

Figure 3. An example of applying model-based SS in the analysis of a forest recording. The input data are a spectrogram generated from a field recording collected in Fonghuanggu Bird and Ecology Park, a broad-leaved evergreen forest located in central Taiwan. The spectrogram was whitened using the 10th percentile of power spectral densities at each frequency band. We manually annotated the presence (black circles) of three types of animal vocalizations and environmental sounds to train a deep NMF model, which has four layers of convolutive NMF. The result shows that the deep NMF model accurately captures the spectral modulation of bird songs (source 1) and two types of insect calls (source 2 and 3).

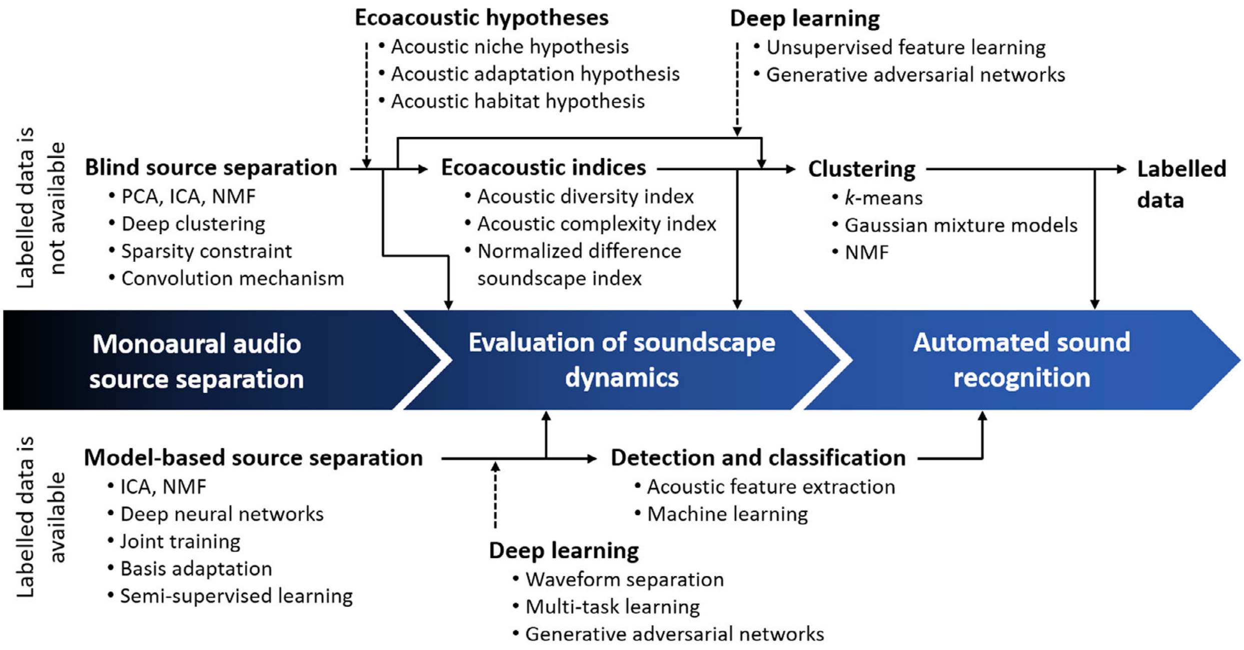
Ecoacoustic hypotheses
• Acoustic niche hypothesis
• Acoustic adaptation hypothesis
• Acoustic habitat hypothesis
Deep learning
• Unsupervised feature learning
• Generative adversarial networks
Labelled data is not available
Blind source separation
• PCA, ICA, NMF
• Deep clustering
• Sparsity constraint
• Convolution mechanism
Ecoacoustic indices
• Acoustic diversity index
• Acoustic complexity index
• Normalized difference soundscape index
Clustering
• k-means
• Gaussian mixture models
• NMF
Labelled data
Monoaural audio source separation
Evaluation of soundscape dynamics
Automated sound recognition
Labelled data is available
Model-based source separation
• ICA, NMF
• Deep neural networks
• Joint training
• Basis adaptation
• Semi-supervised learning
Detection and classification
• Acoustic feature extraction
• Machine learning
Deep learning
• Waveform separation
• Multi-task learning
• Generative adversarial networks

Figure 4. A roadmap to integrate monoaural audio source separation with other techniques of soundscape information retrieval. Dashed lines suggest future integrations with ecoacoustics and informatics to further elevate the performance of blind source separation and model-based source separation.

the results) and subsequently apply supervised training to improve the separation performance. Many wildlife produce calls with time-varying frequency modulation. A convolutive form of machine learning, for example, CNN or convolutive NMF, will be an effective solution to capture source characteristics accurately. We can also enforce a sparsity constraint according to source behaviour, such as calling rate, to increase the opportunity of learning independent sources. However, satisfactory performance will only be achieved when assumptions of the source behaviour are appropriate.

Even though the current technology of SS has many challenges, SS can still be used to improve conventional approaches, such as ecoacoustic indices, clustering, detection and classification, to achieve a versatile soundscape information retrieval (Fig. 4). Future developments of SS can consider integrating ecological hypotheses into the learning procedure. For example, the Acoustic Niche Hypothesis predicts that sounds of species sharing the same acoustic space have evolved such that the acoustic competition in both frequency and time is minimized (Ruppé et al. 2015). Thus, the hypothesis can be integrated with BSS to facilitate the assessment of the diversity of soniferous animals. In model-based SS, labelled data may be applied in multi-task learning to derive better representations that capture invariant properties, and improve the semantic classification of soundscapes (Maurer et al. 2016). Generative adversarial networks, which have been shown to perform well in many generation tasks, may also be a potential direction for audio source separation (Subakan and Smaragdis 2018; Stoller et al. 2018b; Fu et al. 2019).

We have addressed applications of monoaural audio SS techniques; however, many separation techniques are available for microphone and hydrophone arrays. SS in an overdetermined system will also improve the monitoring of soundscape dynamics. While more acoustic sensors are deployed for longer durations to investigate global and regional soundscape changes, soundscape-based ecosystem monitoring nevertheless runs into a problem of scalability (Stowell 2018). As the volume of data grows, with the expectation of real-time processing, in conservation management, the development of efficient frameworks of SS will be critical. Therefore, we offer this study as the first step towards versatile soundscape information retrieval, and we wish to extend an invitation of cross-disciplinary collaboration to researchers in ecoacoustics and informatics.

Acknowledgements

This work was supported by the JSPS KAKENHI, Japan (Grant Number JP18H06491), and the project entitled “Extending Asian Soundscape Monitoring Network and Moving Towards Open Science Platform” (Asi@Connect-17-100).

References

André, M., M. van der Schaar, S. Zaugg, L. Houégnigan, A. M. Sánchez, and J. V. Castell. 2011. Listening to the Deep: live monitoring of ocean noise and cetacean acoustic signals. *Mar. Pollut. Bull.* **63**, 18–26.

Bardeli, R., D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt. 2010. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recogn. Lett.* **31**, 1524–1534.

Bellisario, K. M., and B. C. Pijanowski. 2019. Contributions of MIR to soundscape ecology. Part I: potential methodological synergies. *Ecol. Inform.* **51**, 96–102.

Bellisario, K. M., J. VanSchaik, Z. Zhao, A. Gasc, H. Omrani, and B. C. Pijanowski. 2019a. Contributions of MIR to soundscape ecology. Part 2: spectral timbral analysis for discriminating soundscape components. *Ecol. Inform.* **51**, 1–14.

Bellisario, K. M., T. Broadhead, D. Savage, Z. Zhao, H. Omrani, S. Zhang, et al. 2019b. Contributions of MIR to soundscape ecology. Part 3: tagging and classifying audio features using a multilabeling *k*-nearest neighbor approach. *Ecol. Inform.* **51**, 103–111.

Blumstein, D. T., D. J. Mennill, P. Clemins, L. Girod, K. Yao, G. Patricelli, et al. 2011. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. *J. Appl. Ecol.* **48**, 758–767.

Bohnenstiehl, D. R., R. P. Lyon, O. N. Caretti, S. W. Ricci, and D. B. Eggleston. 2018. Investigating the utility of ecoacoustic metrics in marine soundscapes. *J. Ecoacoustics* **2**, R1156L.

Bryan, N. J., and G. J. Mysore. 2013.Interactive refinement of supervised and semi-supervised sound source separation estimates. Pp.883–887 in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Cano, E., D. FitzGerald, A. Liutkus, M. D. Plumbley, and F.-R. Stoter. 2019. Musical source separation: an introduction. *IEEE Signal Proc. Mag.* **36**, 31–40.

Chen, Z., S. Watanabe, H. Erdogan, and J. R. Hershey. 2015. Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks. Pp.3274–3278in *Interspeech 2015*.

Comon, P. 1994. *Independent component analysis, A new concept?Signal Process.* **36**, 287–314.

Coquereau, L., J. Lossent, J. Grall, and L. Chauvaud. 2017. Marine soundscape shaped by fishing activity. *Roy. Soc. Open Sci.* **4**, 160606.

Du, J., Y. Tu, Y. Xu, L. Dai, and C.-H. Lee. 2014. Speech separation of a target speaker based on deep neural networks. Pp.473–477 in *2014 12th International Conference on Signal Processing (ICSP)*.

Dumyahn, S. L., and B. C. Pijanowski. 2011. Soundscape conservation. *Landscape Ecol.* **26**, 1327–1344.

Eldridge, A., M. Casey, P. Moscoso, and M. Peck. 2016. A new method for ecoacoustics? Toward the extraction and evaluation of ecologically-meaningful soundscape components using sparse coding methods. *PeerJ* **4**, e2108.

Erdogan, H., J. R. Hershey, S. Watanabe, and J. Le Roux. 2015. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. Pp.708–712 in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Ewert, S., and M. B. Sandler. Structured dropout for weak label and multi-instance learning and its application to score-informed source separation.Pp. 2277–2281 in . 2017. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Ey, E., and J. Fischer. 2009. The acoustic adaptation hypothesis: a review of the evidence from birds, anurans and mammals. *Bioacoustics* **19**, 21–48.

Fairbrass, A. J., P. Rennert, C. Williams, H. Titheridge, and K. E. Jones. 2017. Biases of acoustic indices measuring biodiversity in urban areas. *Ecol. Indic.* **83**, 169–177.

Fan, H.-T., J.-W. Hung, X. Lu, S.-S. Wang, and Y. Tsao. 2014. Speech enhancement using segmental nonnegative matrix factorization. Pp.4483–4487 in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Farina, A. 2018. Perspectives in ecoacoustics: a contribution to defining a discipline. *J. Ecoacoustics*, **2**, TRZD5I.

Fu, S.-W., C.-F. Liao, Y. Tsao, and S.-D. Lin. 2019. MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement. ArXiv, 1905.04874.

Fu, Z., G. Lu, K. M. Ting, and D. Zhang. 2011. A survey of audio-based music classification and annotation. *IEEE T. Multimedia* **13**(2), 303–319.

Fu, S.-W., Y. Tsao, X. Lu, and H. Kawai. 2017. Raw waveform-based speech enhancement by fully convolutional networks. Pp.6–12 in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

Fu, S.-W., T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai. 2018. End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. *IEEE T. Audio Speech* **26**, 1570–1584.

Gillespie, D., D. K. Mellinger, J. Gordon, D. McLaren, P. Redmond, R. McHugh, et al. 2009. PAMGUARD: semiautomated, open source software for real-time acoustic detection and localization of cetaceans. *J. Acoust. Soc. Am.* **125**, 2547–2547.

Guan, S., T.-H. Lin, L.-S. Chou, J. Vignola, J. Judge, and D. Turo. 2015. Dynamics of soundscape in a shallow water marine environment: a study of the habitat of the Indo-Pacific humpback dolphin.*J. Acoust. Soc. Am.* **137**, 2939–2949.

Hershey, J. R., Z. Chen, J. Le Roux, and S. Watanabe. 2016. *Deep clustering: discriminative embeddings for segmentation and separation*. Signal Processing (ICASSP).

Hinton, G. E., and R. R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507.

Hoyer, P. O. 2004. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res* **5**, 1457–1469.

Huang, P.-S., S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson. 2012. Singing-voice separation from monaural recordings using robust principal component analysis. Pp.57–60 *in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Huang, P.-S., M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. 2015. Joint optimization of masks and deep recurrent neural networks for monaural source separation.. *IEEE T. Audio Speech* **23**, 2136–2147.

Hui, L., M. Cai, C. Guo, L. He, W.-Q. Zhang, and J. Liu. 2015. Convolutional maxout neural networks for speech separation. Pp.24-27 in *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*.

Innami, S., and H. Kasai. 2012. NMF-based environmental sound source separation using time-variant gain features. *Comput. Math. Appl.* **64**, 1333–1342.

Isik, Y., J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey. 2016. Single-channel multi-speaker separation using deep clustering. Pp.545–549 in *Interspeech* 2016.

Jang, G.-J., and T.-W. Lee. 2003. A maximum likelihood approach to single-channel source separation. *J. Mach. Learn. Res.* **4**, 1365–1392.

Kameoka, H., T. Higuchi, M. Tanaka, and L. Li. 2018. Nonnegative matrix factorization with basis clustering using cepstral distance regularization.. *IEEE T. Audio Speech* **26**, 1029–1040.

Karamatlı, E., A. T. Cemgil, and S. Kırbız. 2018. Weak label supervision for monaural source separation using non-negative denoising variational autoencoders. *ArXiv*, 1810.13104.

Kasten, E. P., H. G. Stuart, J. Fox, and W. Joo. 2012. The remote environmental assessment laboratory's acoustic library: an archive for studying soundscape ecology. *Ecol. Inform.* **12**, 50–67.

Kerr, J. T., and M. Ostrovsky. 2003. From space to species: ecological applications for remote sensing. *Trends Ecol. Evol.* **18**, 299–305.

Kingma, D. P., and M. Welling. 2014. *Auto-encoding variational bayes. in*. Learning Representations (ICLR).

Krause, B. L. 1993. The niche hypothesis: a virtual symphony of animal sounds, the origins of musical expression and the health of habitats. *Soundscape Newsl.* **6**, 4–6.

Krause, B., and A. Farina. 2016. Using ecoacoustic methods to survey the impacts of climate change on biodiversity. *Biol. Conserv.* **195**, 245–254.

Kuehne, L. M., B. L. Padgham, and J. D. Olden. 2013. The soundscapes of lakes across an urbanization gradient. *PLoS ONE* **8**, e55661.

Längkvist, M., L. Karlsson, and A. Loutfi. 2014. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recogn. Letters* **42**, 11–24.

Le Roux, J., J. R. Hershey, and F. Weninger. 2015. Deep NMF for speech separation. Pp.66–70 in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* **521**, 436–444.

Lee, D. D., and H. S. Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791.

Lin, T.-H., and Y. Tsao. 2018. Listening to the deep: Exploring marine soundscape variability by information retrieval techniques. Pp.1–6 in *2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO)*.

Lin, T.-H., L.-S. Chou, T. Akamatsu, H.-C. Chan, and C.-F. Chen. 2013. An automatic detection algorithm for extracting the representative frequency of cetacean tonal sounds. *J. Acoust. Soc. Am.* **134**, 2477–2485.

Lin, T.-H., S.-H. Fang, and Y. Tsao. 2017a. Improving biodiversity assessment via unsupervised separation of biological sounds from long-duration recordings. *Sci. Rep.* **7**, 4547.

Lin, T.-H., Y. Tsao, Y.-H. Wang, H.-W. Yen, and S.-S. Lu. 2017b. Computing biodiversity change via a soundscape monitoring network. Pp.128–133 in *2017 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*.

Lin, T.-H., Y. Tsao, and T. Akamatsu. 2018. Comparison of passive acoustic soniferous fish monitoring with supervised and unsupervised approaches. *J. Acoust. Soc. Am.* **143**, EL278.

Lin, T.-H., H.-T. Yang, J.-M. Huang, C.-J. Yao, Y.-S. Lien, P.-J. Wang, et al. 2019. Evaluating changes in the marine soundscape of an offshore wind farm via the machine learning-based source separation. Pp.1–6 in *2019 IEEE Underwater Technology (UT)*.

Lostanlen, V., K. Palmer, E. Knight, C. Clark, H. Klinck, A. Farnsworth, et al. 2019. Long-distance detection of bioacoustic events with per-channel energy normalization. *ArXiv* **1911**, 00417.

Lu, X., Y. Tsao, S. Matsuda, and C. Hori. 2013. Speech enhancement based on deep denoising autoencoder. Pp.436–440in *Interspeech* 2013.

Maurer, A., M. Pontil, and B. Romera-Paredes. 2016. The benefit of multitask representation learning. *J. Mach. Learn. Res.* **17**, 1–32.

Menze, S., D. P. Zitterbart, I. van Opzeeland, and O. Boebel. 2017. The influence of sea ice, wind speed and marine mammals on Southern Ocean ambient sound. *Roy. Soc. Open Sci.* **4**, 160370.

Molla, M. K. I., K. Hirose, and N. Minematsu. 2008. The robustness and applicability of audio source separation from single mixtures. *Acoust. Aust.* **36**, 2–55.

Monczak, A., C. Mueller, M. E. Miller, Y. Ji, S. A. Borgianini, and E. W. Montie. 2019. Sound patterns of snapping shrimp, fish, and dolphins in an estuarine soundscape of the southeastern USA. *Mar. Ecol. Prog. Ser.* **609**, 49–68.

Mullet, T. C., A. Farina, and S. H. Gage. 2017. The acoustic habitat hypothesis: An ecoacoustics perspective on species habitat selection. *Biosemiotics* **10**, 319–336.

Nugraha, A. A., A. Liutkus, and E. Vincent. 2016. Multichannel audio source separation with deep neural networks. *IEEE T. Audio Speech* **24**, 1652–1664.

O'Grady, P., and B. Pearlmutter. 2006. Convolutive non-negative matrix factorisation with a sparseness constraint. Pp. 427-432 in *2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*.

Phillips, Y. F., M. Towsey, and P. Roe. 2018. Revealing the ecological content of long-duration audio-recordings of the environment through clustering and visualisation. *PLoS ONE* **13**, e0193345.

Pieretti, N., A. Farina, and D. Morri. 2011. A new methodology to infer the singing activity of an avian community: the Acoustic Complexity Index (ACI). *Ecol. Indic.* **11**, 868–873.

Pijanowski, B. C., L. J. Villanueva-Rivera, S. L. Dumyahn, A. Farina, B. L. Krause, B. M. Napoletano, et al. 2011. Soundscape ecology: the science of sound in the landscape. *Bioscience* **61**, 203–216.

Ross, S. R. P.-J., N. R. Friedman, K. L. Dudley, M. Yoshimura, T. Yoshida, and E. P. Economo. 2018. Listening to ecosystems: data-rich acoustic monitoring through landscape-scale sensor networks. *Ecol. Res.* **33**, 135–147.

Ruppé, L., G. Clément, A. Herrel, L. Ballesta, T. Décamps, L. Kéver, et al. 2015. Environmental constraints drive the partitioning of the soundscape in fishes. *P. Natl. Acad. Sci. USA* **112**, 6092–6097.

Saito, K., K. Nakamura, M. Ueta, R. Kurosawa, A. Fujiwara, H. H. Kobayashi, et al. 2015. Utilizing the Cyberforest live sound system with social media to remotely conduct woodland bird censuses in Central Japan. *Ambio* **44**, 572–583.

Smaragdis, P., B. Raj, and M. Shashanka. 2007. Supervised and semi-supervised separation of sounds from single-channel mixtures.Pp.414–421 in *Independent Component Analysis and Signal Separation (ICA'07)*.

Smaragdis, P., C. Fevotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman. 2014. Static and dynamic source separation using nonnegative factorizations: a unified view. *IEEE Signal Proc. Mag.* **31**, 66–75.

Sobieraj, I., Q. Kong, and M. D. Plumbley. 2017. Masked non-negative matrix factorization for bird detection using weakly labeled data. Pp.1769–1773 in *2017 25th European Signal Processing Conference (EUSIPCO)*.

Stoller, D., S. Ewert, and S. Dixon. 2018a. Wave-u-net: a multiscale neural network for end-to-end source separation. Pp.334–340 in *19th International Society for Music Information Retrieval Conference (ISMIR)*.

Stoller, D., S. Ewert, and S. Dixon. 2018b. Adversarial semi-supervised audio source separation applied to singing voice extraction.Pp. 2391–2395 in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Stöter, F.-R., A. Liutkus, and N. Ito. 2018. The 2018 signal separation evaluation campaign.Pp. 293–305 in *International Conference on Latent Variable Analysis and Signal Separation*.

Stowell, D. 2018. Computational bioacoustic scene analysis.Pp.303-333 in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. D. Plumbley and D. Ellis, Eds. Berlin, Germany: Springer.

Stowell, D., and R. E. Turner. 2015. Denoising without access to clean data using a partitioned autoencoder. *ArXiv* **1509**, 05982.

Subakan, Y. C., and P. Smaragdis. 2018. Generative adversarial source separation. Pp.26–30 in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Sueur, J., and A. Farina. 2015. Ecoacoustics: the ecological investigation and interpretation of environmental sound. *Biosemiotics* **8**, 493–502.

Sueur, J., S. Pavoine, O. Hamerlynck, and S. Duvail. 2008. Rapid acoustic survey for biodiversity appraisal. *PLoS ONE* **3**, e4065.

Sueur, J., A. Farina, A. Gasc, N. Pieretti, and S. Pavoine. 2014. Acoustic indices for biodiversity assessment and landscape investigation. *Acta Acust. united Ac.* **100**, 772–781.

Suzuki, R., S. Matsubayashi, R. W. Hedley, K. Nakadai, and H. G. Okuno. 2017. HARKBird: exploring acoustic interactions in bird communities using a microphone array. *J. Robot. Mechatron* **27**, 213–223.

Tu, Y., J. Du, Y. Xu, L. Dai, and C.-H. Lee. 2014. *Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers*. Spoken Language Processing.

Uhlich, S., F. Giron, and Y. Mitsufuji. 2015. Deep neural network based instrument extraction from music. Pp. 2135–2139 in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Ulloa, J. S., T. Aubin, D. Llusia, C. Bouveyron, and J. Sueur. 2018. Estimating animal acoustic diversity in tropical environments using unsupervised multiresolution analysis. *Ecol. Indic.* **90**, 346–355.

Vaseghi, S. V. 2008. Advanced digital signal processing and noise reduction (Fourth Edition). John Wiley & Sons.

Villanueva-Rivera, L. J., B. C. Pijanowski, J. Doucette, and B. Pekin. 2011. A primer of acoustic analysis for landscape ecologists. *Landscape Ecol.* **26**, 1233–1246.

Virtanen, T. 2006. Unsupervised learning methods for source separation in monaural music signals. Pp. 267–296 in A.

Klapuri, M. Davy, eds. *Signal Processing Methods for Music Transcription*. Springer, New York.

Wang, D., and J. Chen. 2017. Supervised speech separation based on deep learning: an overview. *IEEE T. Audio Speech* **26**, 1702–1726.

Wang, Z., and F. Sha. 2014. Discriminative non-negative matrix factorization for single-channel speech separation. Pp.3749–3753 in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Wang, Y., and D. Wang. 2013. Towards scaling up classification-based speech separation. *IEEE T. Audio Speech* **21**, 1381–1390.

Weninger, F., J. L. Roux, J. R. Hershey, and S. Watanabe. 2014. Discriminative NMF and its application to single-channel source separation. Pp.865–869 in *Interspeech 2014*.

Wold, E., T. Blum, D. Keislar, and J. Wheaten. 1996. Content-based classification, search, and retrieval of audio. *IEEE Multimedia* **3**, 27–36.

Wu, C.-W., and A. Lerch. 2015. Drum transcription using partially fixed non-negative matrix factorization. Pp.1281-1285 in *2015 23rd European Signal Processing Conference (EUSIPCO)*.

Xie, J., M. Towsey, J. Zhang, X. Dong, and P. Roe. 2015. Application of image processing techniques for frog call classification. Pp.4190–4194 in *2015 IEEE International Conference on Image Processing (ICIP)*.

Xu, Y., J. Du, L.-R. Dai, and C.-H. Lee. 2015. A regression approach to speech enhancement based on deep neural networks. *IEEE T. Audio Speech* **23**, 7–19.