Perspectives

Emmanuel Vincent, Tuomas Virtanen, Sharon Gannot

19
Perspectives

Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot

Source separation and speech enhancement research has made dramatic progress in the last 30 years. It is now a mainstream topic in speech and audio processing, with hundreds of papers published every year. Separation and enhancement performance have greatly improved and successful commercial applications are increasingly being deployed. This chapter provides an overview of research and development perspectives in the field. We do not attempt to cover all perspectives currently under discussion in the community. Instead, we focus on five directions in which we believe major progress is still possible: getting the most out of deep learning, exploiting phase relationships across time-frequency bins, improving the estimation accuracy of multichannel parameters, addressing scenarios involving multiple microphone arrays or other sensors, and accelerating industry transfer. These five directions are covered in Sections 19.1, 19.2, 19.3, 19.4, and 19.5, respectively.

19.1
Advancing deep learning

In just a few years, deep learning has emerged as a major paradigm for source separation and speech enhancement. Deep neural networks (DNNs) can model the complex characteristics of audio sources by making efficient use of large amounts (typically, hours) of training data. They perform well on mixtures involving similar conditions to those in the training set and they are surprisingly robust to unseen conditions (Vincent *et al.*, 2017; Kolbæk *et al.*, 2017), provided that the training set is sufficiently large and diverse. Several research directions are currently under study to get the best out of this new paradigm.

19.1.1
DNN design choices

The first obvious direction is to tune the DNN architecture to the task at hand. Several architectures, namely multilayer perceptron, deep recurrent neural network (DRNN),

long short-term memory (LSTM), bidirectional LSTM, convolutional neural network, and nonnegative DNN have already been covered in Section 7.3.2. Recently, a new DRNN-like architecture known as the *deep stacking network* was successfully employed (Zhang *et al.*, 2016). This architecture concatenates the outputs in the previous time frame with the inputs in the current frame. It is motivated by the fact that iteratively applying a DNN to the outputs of the previous iteration improves performance (Nugraha *et al.*, 2016a), but it avoids multiple passes over the test data. Another architecture recently proposed by Chazan *et al.* (2016) combines a generative Gaussian mixture model (GMM) and a discriminative DNN in a hybrid approach. The DNN is used to estimate the posterior phoneme probability in each time frame, and a soft time-frequency mask is derived by modeling each phoneme as a single Gaussian. New architectures are invented every year in the fields of automatic speech and image recognition, e.g., (Zagoruyko and Komodakis, 2016), and it is only a matter of time before they are adapted and applied to source separation and speech enhancement. Fusing the outputs of multiple architectures is also beneficial. The optimal fusion weights can be learned using a DNN, as shown by Jaureguiberry *et al.* (2016).

Another interesting research direction concerns the design of the training set. The most common approach today is to generate simulated training data by convolving target and interference signals with real or simulated acoustic impulse responses and mixing them together. It is generally believed that the larger the amount of training data, the more diverse, and the closer to the test conditions, the better the separation or enhancement performance. This has led to several data augmentation approaches to expand the size and coverage of the training set and reduce the mismatch with the test set. Yet, surprisingly, Heymann *et al.* (2016) obtained similar performance by training on time-frequency masks generated by thresholding target short-time Fourier transform (STFT) coefficients (without using any interference signal in training), while Vincent *et al.* (2017) found that training on mismatched noise conditions can outperform matched or multicondition training. This calls for a more principled approach to designing the training set. The algorithm of Sivasankaran *et al.* (2017), which weights the training samples so as to maximize performance on the validation set, is a first step in this direction.

The cost function used for training also has a significant impact. Various cost functions have been reviewed in Sections 7.3.3 and 7.3.4, namely cross-entropy, mean square error (MSE), phase-sensitive cost, Kullback-Leibler (KL) divergence, and Itakura-Saito (IS) divergence. The studies cited in these sections reported better performance for MSE and the phase-sensitive cost. More recently, however, Nugraha *et al.* (2016a) found KL to perform best in both single- and multichannel scenarios. This calls for more research on the choice of the cost function depending on the scenario and other DNN design choices. Taking psychoacoustics into account is also a promising direction, as recently explored by Shivakumar and Georgiou (2016).

Finally, the use of DNNs also impacts the signal processing steps involved in the overall separation or enhancement system. For instance, Nugraha *et al.* (2016b) found that, when the source power spectra are estimated by a DNN, the conventional expectation-maximization (EM) update rule for spatial covariance matrices (14.79)

is outperformed by a temporally weighted rule. Deeper investigation of the interplay between DNN and signal processing is thus required in order to get the best out of hybrid systems involving both DNN and signal processing steps.

19.1.2 End-to-end approaches

End-to-end DNN-based approaches attempt to address the interplay between DNN and signal processing by getting rid of signal processing entirely and developing purely DNN-based systems as opposed to hybrid systems. This makes it easier to jointly optimize all processing steps in a DNN framework (see Section 17.4.2). End-to-end DNNs operate in the time domain or in the complex-valued STFT domain (Li *et al.*, 2016), which enables them to exploit phase differences between neighboring time-frequency bins (see Section 19.2.3 below). Current end-to-end DNN architectures that integrate target localization and beamforming perform only marginally better than delay-and-sum (DS) beamforming, and lie significantly behind the conventional signal processing-based beamformers derived from DNN-based masks reviewed in Section 12.4 (Xiao *et al.*, 2016). Nevertheless, it is a widespread belief in the deep learning community that they will soon outperform other approaches. Progress might come from *generative DNNs* such as Wavenet (van den Oord *et al.*, 2016) which are now used to synthesize time-domain audio signals and could soon be used to generate the source signals that best match a given mixture signal. As a matter of fact, synthesis-based speech enhancement has recently started being investigated (Nickel *et al.*, 2013; Kato and Milner, 2016).

19.1.3 Unsupervised separation

DNNs are typically trained in a supervised fashion in order to discriminate a certain class of sounds, e.g., speech vs. noise, foreground vs. background speech, male vs. female speech, or a specific speaker vs. others. To do so, the number of sources must be known and the order of the sources must be fixed, i.e., training is permutation-dependent. This implies that separating two foreground speakers of the same gender is infeasible unless their identity is known and training data are available for at least one of them, a situation which arises in certain scenarios only. *Deep clustering* algorithms inspired by the spectral clustering algorithms in Section 7.1.3 have recently overcome this limitation for single-channel mixtures.

Hershey *et al.* (2016) proposed to apply a DRNN $g_{\mathcal{Z}}$ to the log-magnitude spectrogram $\log|\mathbf{X}| = [\log|x(n,f)|]_{fn}$ to extract a unit-norm *embedding*, i.e., a feature vector $\mathbf{y}(n,f)$ of arbitrary dimension K for each time-frequency bin. The output $g_{\mathcal{Z}}(\log|\mathbf{X}|)(n)$ of the DRNN in a given time frame n consists of the concatenation

of the embeddings for all frequency bins in that time frame:

$$\begin{bmatrix} \mathbf{y}(n,0) \\ \vdots \\ \mathbf{y}(n,F-1) \end{bmatrix} = g_{\mathcal{Z}}(\log|\mathbf{X}|)(n). \tag{19.1}$$

The key is to train $g_{\mathcal{Z}}$ such that each embedding characterizes the dominant source in the corresponding time-frequency bin. Separation can then be achieved by clustering these embeddings, so that time-frequency bins dominated by the same source are clustered together. The optimal assignment of time-frequency bins to sources is given by the $FN \times J$ indicator matrix $\mathbf{O} = [o_j(n,f)]_{fnj}$ where

$$o_j(n,f) = \begin{cases} 1 & \text{if source } j \text{ dominates in time-frequency bin } (n,f) \\ 0 & \text{otherwise.} \end{cases} \tag{19.2}$$

All embeddings are stacked into an $FN \times K$ matrix $\mathbf{Y} = [\mathbf{y}^T(n,f)]_{fn}$ and training is achieved by minimizing

$$\mathcal{C}^{\text{PI}}(\mathcal{Z}) = \|\mathbf{Y}\mathbf{Y}^T - \mathbf{O}\mathbf{O}^T\|_2^2. \tag{19.3}$$

This cost is permutation-invariant: indeed, the binary affinity matrix $\mathbf{O}\mathbf{O}^T$ is such that $(\mathbf{O}\mathbf{O}^T)_{fn,f'n'} = 1$ if (n,f) and (n',f') belong to the same cluster and $(\mathbf{O}\mathbf{O}^T)_{fn,f'n'} = 0$ otherwise, and it is invariant to reordering of the sources.

Once the embeddings have been obtained, the sources are separated by soft time-frequency masking. The masks $w_j(n,f)$ are computed by a soft k-means algorithm, which alternately updates the masks and the centroid embedding $\bar{\mathbf{y}}_j$ of each source:

$$w_j(n,f) = \frac{e^{-\alpha\|\mathbf{y}(n,f)-\bar{\mathbf{y}}_j\|_2^2}}{\sum_{j'=1}^{J} e^{-\alpha\|\mathbf{y}(n,f)-\bar{\mathbf{y}}_{j'}\|_2^2}} \tag{19.4}$$

$$\bar{\mathbf{y}}_j = \frac{1}{\sum_{fn} w_j(n,f)} \sum_{fn} w_j(n,f)\mathbf{y}(n,f). \tag{19.5}$$

The parameter α controls the "hardness" of the clustering. In practice, only nonsilent time-frequency bins are taken into account in (19.3) and (19.5).

One limitation of this approach is that the training criterion (19.3) is not directly linked with the source separation performance. To address this, Isik *et al.* (2016) introduce a second DRNN $g_{\mathcal{Z}'}$ that takes as input the mixture amplitude spectrogram $|\mathbf{X}|$ and the amplitude spectrogram of a given source $|\mathbf{C}_j| = [w_j(n,f)|x(n,f)|]_{fn}$ estimated by the above clustering procedure and outputs an improved estimate of $|\mathbf{C}_j|$. An improved soft mask is then computed from these improved estimates and used to obtain the final source estimates. The iterations of the k-means algorithm are unfolded and trained along with this second DRNN according to the signal reconstruction cost (7.23).

The results in Table 19.1 show that this approach can separate mixtures of unknown speakers remarkably better than a DRNN trained to separate the foreground speaker.

An alternative permutation-invariant DNN training approach which does not require intermediate embeddings was proposed by Yu *et al.* (2016). These approaches open up new perspectives for research in DNN-based source separation.

Method	SDR (dB)
Hu and Wang (2013)	3.1
DRNN foreground	1.2
Deep clustering $g_{\mathcal{Z}}$	10.3
Deep clustering $g_{\mathcal{Z}} + g_{\mathcal{Z}'}$	10.8

Table 19.1 Average signal-to-distortion ratio (SDR) achieved by the computational auditory scene analysis (CASA) method of Hu and Wang (2013), a DRNN trained to separate the foreground speaker, and two variants of deep clustering for the separation of mixtures of two speakers with all gender combinations at random signal-to-noise ratios (SNRs) between 0 and 10 dB (Hershey *et al.*, 2016; Isik *et al.*, 2016). The test speakers are not in the training set.

19.2 Exploiting phase relationships

The filters introduced in Chapters 5 and 10 for separation and enhancement exhibit two fundamental limitations. First, due to the narrowband approximation (2.11), they operate in each time-frequency bin independently and can generate cyclic convolution artifacts. Second, except for magnitude spectral subtraction, they rely on the assumption that phase is uniformly distributed, which translates into modeling the target and interference STFT coefficients as zero-mean. As a result, they exploit the magnitude spectrum of each source and the interchannel phase difference (IPD), but not the phase spectrum of each individual channel[1)]. This assumption is motivated by the fact that the phase spectrum appears to be uniformly distributed when wrapped to its principal value in $[0, 2\pi)$. Yet, it does have some structure which can be exploited to estimate phase-aware filters or interframe/interband filters, as recently overviewed by Gerkmann *et al.* (2015) and Mowlaee *et al.* (2016).

19.2.1 Phase reconstruction and joint phase-magnitude estimation

Griffin and Lim (1984) proposed one of the first algorithms to reconstruct the phase from a magnitude-only spectrum. They formulated this problem as finding the time-domain signal whose magnitude STFT is closest to this magnitude spectrum. Starting from a single-channel zero-phase spectrum $c(n, f)$, they iteratively update it as

1) To be precise, the phase of the filtered signal depends on the phase of the individual channels, but the coefficients of the filter depend only on the IPD.

follows:

$$c(n,f) \leftarrow |c(n,f)| \angle \mathrm{STFT}(\mathrm{iSTFT}(c))(n,f) \tag{19.6}$$

where $\mathrm{iSTFT}(\cdot)$ denotes the inverse STFT. This update can be more efficiently implemented in the STFT domain (Le Roux *et al.*, 2010) and it can be shown to result in a local minimum of the above optimization problem. This algorithm exploits the so-called *consistency* of complex-valued spectra (Le Roux *et al.*, 2010), that is the fact that the STFTs of (real-valued) time-domain signals lie in a lower-dimensional linear subspace of the space of $F \times N$ complex-valued matrices, with F the number of frequency bins and N the number of time frames. In other words, magnitude and phase spectra are deterministically related to each other across frequency bins and, in the case when the STFT analysis windows overlap, also across frames[2)]. In a source separation framework, while it can be employed to reestimate the phase of each estimated source spatial image $\widehat{c}_j(n,f)$ independently of the others, it does not exploit the available complex-valued mixture spectrum $x(n,f)$. Gunawan and Sen (2010) and Sturmel and Daudet (2012) described alternative algorithms to jointly reconstruct the phase spectrum of all sources such that the sum of all $\widehat{c}_j(n,f)$ is closest to $x(n,f)$. Kameoka *et al.* (2009) combined this idea with a nonnegative matrix factorization (NMF) model for the magnitudes. The above algorithms only modify the phase spectra of the sources, therefore they can improve the separation or enhancement performance only when the magnitude spectra have been accurately modeled or estimated.

To improve performance in practical scenarios where estimating the magnitude spectra is difficult, joint estimation of the magnitude and phase spectra is necessary. Le Roux and Vincent (2013) proposed a consistent Wiener filtering algorithm that iteratively estimates the magnitude and phase spectra of all sources. Assuming that the source STFT coefficients are zero-mean complex Gaussian, this is achieved by maximizing the logarithm of the posterior (5.35) under either a hard consistency constraint $c_j(n,f) = \mathrm{STFT}(\mathrm{iSTFT}(c_j))(n,f)$ for all n, f, or a soft penalty term $\sum_{nf} |c_j(n,f) - \mathrm{STFT}(\mathrm{iSTFT}(c_j))(n,f)|^2$. Mowlaee and Saeidi (2013) introduced an alternative algorithm based on a phase-aware magnitude estimator. Starting from an initial phase estimate, they estimate the magnitude via this estimator, update the phase via (19.6), then reestimate the magnitude from the updated phase, and so on. Both algorithms require several iterations to converge. More recently, Gerkmann (2014) designed a noniterative joint minimum mean square error (MMSE) estimator of magnitude and phase spectra. The resulting magnitude estimate achieves a tradeoff between phase-blind and phase-aware magnitude estimation, while the resulting phase achieves a tradeoff between the noisy phase and the initial phase estimate.

2) When the STFT analysis windows do not overlap, magnitude and phase spectra are not deterministically related to each other anymore across frames since the subspace generated by the STFTs of time-domain signals become equal to the Cartesian product of the subspaces generated by individual frames. However, they remain statistically related to each other as explained in Sections 19.2.2 and 19.2.3.

19.2.2 Interframe and interband filtering

So far, we have only addressed the structure of phases due to the consistency property. Yet, phases typically exhibit additional structure that is unveiled by considering phase differences between successive time frames or frequency bands. Figure 19.1 depicts the derivative of phase with respect to time, known as the *instantaneous frequency* (Stark and Paliwal, 2008), and its negative derivative with respect to frequency, known as the *group delay* (Yegnanarayana and Murthy, 1992). For robust computation of these quantities, see Mowlaee *et al.* (2016). Both representations reveal horizontal or vertical structures due to the periodic or transient nature of sounds.

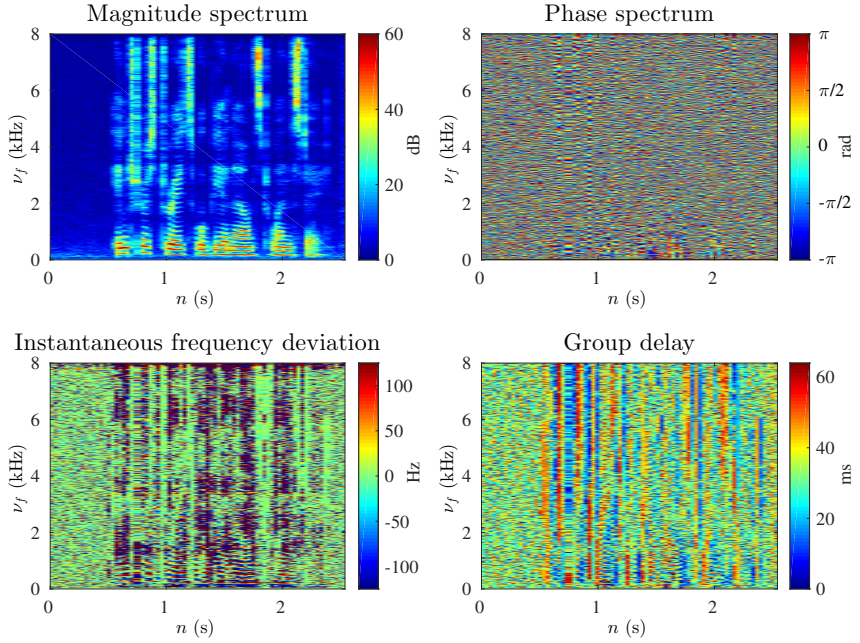

Figure 19.1 Short-term magnitude spectrum and various representations of the phase spectrum of a speech signal for an STFT analysis window size of 64 ms. For easier visualization, the deviation of the instantaneous frequency from the center frequency of each band is shown rather than the instantaneous frequency itself.

Designing *interframe and/or interband filters* that exploit these phase differences can improve single-channel and multichannel enhancement performance. Interframe minimum variance distortionless response (MVDR) beamformers and multichannel

Wiener filters (MWFs) can be designed by stacking N' successive frames

$$\bar{\mathbf{x}}(n,f) = \begin{bmatrix} \mathbf{x}(n,f) \\ \mathbf{x}(n-1,f) \\ \vdots \\ \mathbf{x}(n-N'+1,f) \end{bmatrix} \tag{19.7}$$

and applying the computations in Section 10.4 to $\bar{\mathbf{x}}(n,f)$ instead of $\mathbf{x}(n,f)$ (Avargel and Cohen, 2008; Talmon *et al.*, 2009; Huang and Benesty, 2012; Schasse and Martin, 2014; Fischer and Gerkmann, 2016), where each frame acts as an additional input channel. This implies estimating the interframe covariance matrix of target and interfering sources. Attias (2003) and Kameoka *et al.* (2010) estimated multichannel interframe filters using EM or variational Bayesian (VB) inference instead (see Table 14.1 for other examples). Interband filters can be obtained in a similar way by stacking successive frequency bins (Avargel and Cohen, 2008; Talmon *et al.*, 2009; Huang *et al.*, 2014). Linear prediction-based reverberation cancellation techniques (see Section 15.2) are also examples of single- or multichannel interframe filters. In addition to better exploiting the available phase information, another important attribute of these filters is that they can potentially overcome the circular convolution artifacts inherent to the narrowband approximation via subband filtering (2.10) or STFT-domain filtering (2.9).

19.2.3 Phase models

While they can greatly improve performance in theory, interframe/interband filters are hard to estimate in practice due to the larger number of parameters involved. For instance, the number of entries of interframe covariance matrices grows quadratically with the number N' of stacked frames. To circumvent this issue, Fischer and Gerkmann (2016) considered fixed data-independent interframe coherence matrices trained on a wide range of interference signals. Yet, it is clear that data-dependent parameters are required to benefit from the full potential of such filters. This calls for prior models of phase along time and frequency based on the structure of sounds (see Section 2.2.2). Mowlaee and Saeidi (2013) exploited the fact that the group delay of harmonic sounds is minimum at the harmonics (Yegnanarayana and Murthy, 1992), while Krawczyk and Gerkmann (2014) and Bronson and Depalle (2014) reconstructed the phase of each harmonic assuming a sinusoidal model. Magron *et al.* (2015) used the repetition of musical notes to estimate the phases. Badeau (2011) proposed a probabilistic extension of NMF involving interframe and interband filters that can model both structured or random phases. End-to-end DNNs (see Section 19.1.2) provide a new take on this issue. These studies can be seen as first steps in the direction of better phase modeling. Their improvement and their combination with interframe and interband filtering hold great promise.

19.3 Advancing multichannel processing

With the advent of advanced spectral models such as NMF and DNN, accurately estimating the source power spectra or the source presence probabilities is now possible by jointly exploiting these models and the observed mixture. By contrast, accurately estimating their spatial parameters, e.g., relative acoustic transfer functions (RTFs) or spatial covariance matrices, still remains difficult today. Most methods do not rely on any prior constraint over the spatial parameters but on the observed mixture STFT coefficients only. They can provide accurate spatial parameter estimates only when the source power spectra or the source presence probabilities have been accurately estimated in the first place and the number of time frames is large enough. The difficulty is increased when the sources or the microphones are moving, since the spatial parameters vary for each time frame and statistics cannot be robustly computed from a single frame. We describe below two research directions towards designing more constrained spatial models.

19.3.1 Dealing with moving sources and microphones

Methods designed for moving sources and microphones fall into three categories. The first category of methods simply track the spatial parameters over time using an online learning method (see Section 19.5.1 below). This approach was popular in the early days of sparsity-based separation (Rickard *et al.*, 2001; Lösch and Yang, 2009) and frequency domain independent component analysis (FD-ICA) (Mukai *et al.*, 2003; Wehr *et al.*, 2007), and it is still today for beamforming (Affes and Grenier, 1997; Markovich-Golan *et al.*, 2010). The activity pattern of the sources, i.e., sources appearing or disappearing, can also be tracked by looking at the stability of the parameter estimates over time (Markovich-Golan *et al.*, 2010). Although the estimated parameters are time-varying, the amount of variation over time cannot be easily controlled. Sliding block methods estimate the parameters in a given time frame from a small number of neighboring frames without any prior constraint, which intrinsically limits their accuracy. Exponential decay averaging methods can control the amount of variation over time by setting the decay factor, however the relationship between the actual temporal dynamics of the data and the optimal decay factor is not trivial.

A second category of methods track the spatial parameters over time by explicitly modeling their temporal dynamics. Duong *et al.* (2011) set a continuous Markov chain prior on the time-varying spatial covariance matrix of each source $\mathbf{R}_j(n,f)$

$$\mathbf{R}_j(n,f) \sim \mathcal{IW}(\mathbf{R}_j(n,f) \mid (m-I)\mathbf{R}_j(n-1,f), m) \tag{19.8}$$

where $\mathcal{IW}(. \mid \mathbf{\Psi}, m)$ is the *inverse Wishart* distribution over positive definite matrices with inverse scale matrix $\mathbf{\Psi}$ and m degrees of freedom and I is the number of channels. This prior is such that the mean of $\mathbf{R}_j(n,f)$ is equal to $\mathbf{R}_j(n-1,f)$ and the parameter m controls the deviation from the mean. The spatial covariance matrices can then be estimated in the maximum a posteriori (MAP) sense by EM,

where the M-step involves solving a set of quadratic matrix equations. Kounades-Bastian *et al.* (2016) concatenated the columns of the $I \times J$ time-varying mixing matrix $\mathbf{A}(n, f)$ into an $IJ \times 1$ vector $\bar{\mathbf{a}}(n, f)$ and set a Gaussian continuity prior instead:

$$\bar{\mathbf{a}}(n, f) \sim \mathcal{N}_{\mathrm{c}}(\bar{\mathbf{a}}(n, f) \mid \bar{\mathbf{a}}(n-1, f), \Sigma_{\bar{\mathbf{a}}}(f)). \tag{19.9}$$

They employed a Kalman smoother inside a VB algorithm to estimate the parameters. Such continuity priors help estimating time-varying parameters at low frequencies. Their effectiveness is limited at high frequencies, due to fact that the spatial parameters vary much more quickly. Indeed, a small change in the direction of arrival (DOA) results in a large IPD at high frequencies. Also, they do not easily handle appearing or disappearing sources due to the ensuing change of dimension.

The last category of methods rely on time-varying DOAs and activity patterns estimated by a source localization method (see Chapter 4) in order, e.g., to build a soft time-frequency mask (Pertilä, 2013) or to adapt a beamformer (Madhu and Martin, 2011; Thiergart *et al.*, 2014). These methods usually employ more sophisticated temporal dynamic models involving, e.g., "birth and death" processes and tracking the speed and acceleration of the sources in addition to their position. They allow faster tracking at high frequencies, but they are sensitive to localization errors. Also, they typically exploit the estimated DOAs only, and do not attempt to estimate the deviations of the spatial parameters from their theoretical free-field values. The geometrical constraints reviewed in Chapter 3, that provide a prior model for the deviation of the RTF from the relative steering vector (3.15) and that of the spatial covariance matrix from its mean (3.21), could be used to improve estimation. The integration of DOA-based methods with advanced spectral models and continuity-based methods is also a promising research direction. A first attempt in this direction was made by Higuchi *et al.* (2014).

19.3.2 Manifold learning

The models in Section 19.3.1 are valid for any recording room. A complementary research trend aims to learn a constrained model for the spatial parameters in a specific room. The problem can be formulated as follows: given a set of acoustic transfer functions sampled for a finite number of source and microphone positions, learn the acoustic properties of the room so as to predict the acoustic transfer function for any other source and microphone position. If the acoustic transfer function could be accurately predicted, source localization and spatial parameter estimation would be unified as a single problem, that would be much easier to solve.

The set of acoustic transfer functions in a given room forms a *manifold*. Although acoustic transfer functions are high-dimensional, they live on a small-dimensional nonlinear subspace parameterized by the positions and the orientations of the source and the microphone. This subspace is continuous: nearby source and the microphone positions result in similar acoustic transfer functions. This property extends to RTFs, as illustrated in Fig. 19.2. Once again, the continuity is stronger at low frequencies.

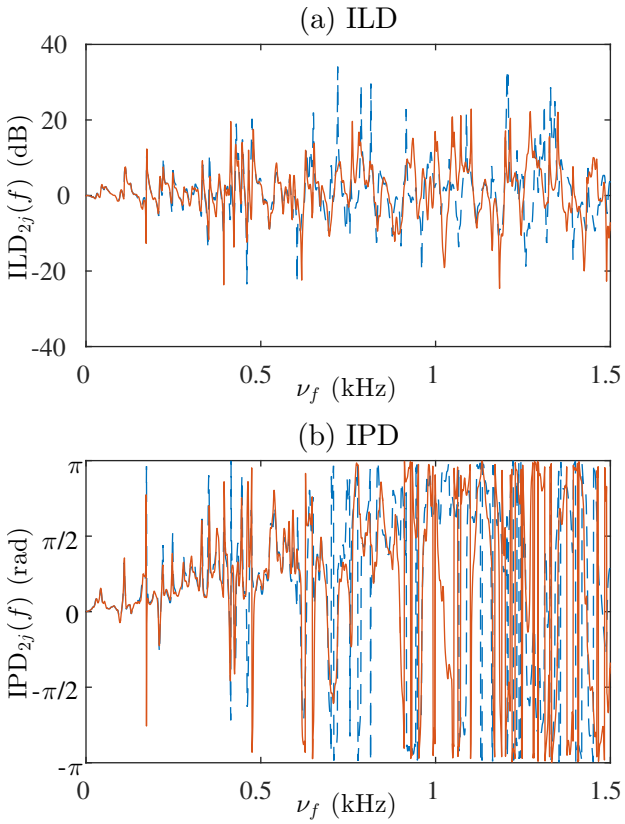

Figure 19.2 Interchannel level difference (ILD) and IPD for two different source positions $j = 1$ (plain red curve) and $j = 2$ (dashed blue curve) 10 cm apart from each other at 1.70 m distance from the microphone pair. The source DOAs are 10° and 13°, respectively. The room size is $8.00 \times 5.00 \times 3.10$ m, the reverberation time is 230 ms, and the microphone distance is 15 cm.

A series of studies have attempted to model this manifold. Koldovský *et al.* (2013) predicted the RTF in a given source position by sparse interpolation of RTFs recorded at nearby positions. Mignot *et al.* (2014) showed that accurate interpolation can be achieved from few samples at low frequencies using a compressed sensing framework that exploits the modal properties of the room, i.e., the spatial frequencies related to the room dimensions. Deleforge *et al.* (2015) introduced a probabilistic piecewise affine mapping model that partitions the space of position and orientation coordinates into regions via a GMM and approximates the RTF within each region as a linear (affine) function of the coordinates, that is similar to the tangent of the manifold. They derived an EM algorithm for jointly learning the GMM and the linear functions from RTF samples. Wang *et al.* (in preparation) investigated the existence of global dependencies beyond the local tangent structure using a DNN with rectified

linear unit activations, which results in a locally linear mapping whose parameters are tied across regions. This DNN outperformed probabilistic piecewise affine mapping and conventional linear interpolation for interpolation distances of 5 cm and beyond, especially at medium to high frequencies, while conventional linear interpolation performed best for shorter distances. Finally, Laufer-Goldshtein *et al.* (2016) demonstrated the limitations of linear approaches to infer physical adjacencies. They defined the diffusion distance related to the geodesic distance on the manifold and demonstrated its ability to arrange the samples according to their DOA and to achieve accurate source localization. This distance also combines local and global properties of the manifold.

While these studies have achieved some success in modeling the manifold from a theoretical point of view and in improving the source localization accuracy, their application to practical source separation and speech enhancement scenarios remains an open issue. Talmon and Gannot (2013) reported a preliminary study of applying these concepts to compute the blocking matrix of a generalized sidelobe canceler (GSC). Deleforge *et al.* (2015) proposed a VB algorithm for joint source localization and separation by soft time-frequency masking, where both the source positions and the index of the dominant source in each time-frequency bin are considered as hidden data. In order to learn the manifold, a certain number of RTF samples must be recorded in the room. This is feasible in scenarios involving robots (Deleforge *et al.*, 2015), but less easily so in other scenarios. Asaei *et al.* (2014) made a first step towards circumventing this requirement by attempting to jointly estimate the room dimensions and the early part of the room impulse response in an unsupervised fashion from the mixture signal alone. Laufer-Goldshtein *et al.* (2016) made another step by proposing a semi-supervised approach that requires only a few RTF samples to be labeled with the source and microphone positions and orientations.

19.4 Addressing multiple-device scenarios

Except for Chapter 18, the separation and enhancement algorithms reviewed in the previous chapters assume that signal acquisition and processing are concentrated in a single device. In many practical scenarios, however, several devices equipped with processing power, wireless communication capabilities, one or more microphones, and possibly other sensors (e.g., accelerometer, camera, laser rangefinder) are available. With current technology, these devices could be connected to form a *wireless acoustic sensor network*. Cellphones, laptops, and tablets, but also webcams, set-top-boxes, televisions, and assistive robots are perfect candidates as nodes (or subarrays) of such networks. Compared with classical arrays, wireless acoustic sensor networks typically comprise more microphones and they cover a wider area. This increases the chance that each target source is close to at least one microphone, hence the potential enhancement performance. However, they raise new challenges regarding signal transmission and processing that should be addressed to fully exploit their potential.

19.4.1 Synchronization and calibration

A first challenge is that the crystal clocks of different devices operate at slightly different frequencies. Therefore, even in the case when the analog-to-digital converters have the same nominal sampling rate, their effective sampling rates differ. The relative deviation ϵ from the nominal sampling rate can be up to $\pm 10^{-4}$. This deviation implies a linear drift ϵt over time, that translates into a phase drift of the STFT coefficients $x_i(n, f)$ that is proportional to ϵ and to the signal frequency ν_f. Figure 19.3 shows that, even for small ϵ, this quickly leads to large phase shifts at high frequencies which preclude the use of multichannel signal processing. The drift can be measured by exchanging time stamps (Schmalenstroeer *et al.*, 2015) or in a blind way by measuring phase shifts between the network nodes (Markovich-Golan *et al.*, 2012a; Miyabe *et al.*, 2015; Wang and Doclo, 2016; Cherkassky and Gannot, 2017). It can then be compensated by resampling the signals in the time domain or applying the opposite phase shift in the STFT domain.

Once it has been compensated, the signals recorded by different devices still have different temporal offsets, which vary slowly over time. In the case when the sources do not move, these offsets do not affect most source separation and speech enhancement methods that do not rely on a prior model of IPDs across nodes but estimate them adaptively from the recorded signals. In the case when the sources are moving, tracking their location over time becomes desirable (see Section 19.3.1) and this requires estimating the offsets. The offsets can be estimated in a similar way as the sampling rate deviation from time stamps (Schmalenstroeer *et al.*, 2015) or phase shifts between the network nodes (Pertilä *et al.*, 2013). The latter approach assumes that the sound scene consists of diffuse noise or many sound sources surrounding the array, since the time difference of arrival (TDOA) due to a single localized source cannot be discriminated from the clock offset. Unfortunately, the above algorithms can estimate the temporal offset up to a standard deviation in the order of 0.1 sample to a few samples in the time domain, which is too large for source localization. Whenever precise localization is required, active self-localization algorithms based on emitting a calibration sound are required (Le and Ono, 2017).

Most of the above algorithms address the estimation of sampling rate mismatch and temporal offset between two nodes only. In the case when the number of nodes is larger than two, one arbitrary node is typically chosen as the master and all other nodes are synced to it. Schmalenstroeer *et al.* (2015) showed how to efficiently synchronize all nodes towards a virtual master node, which represents the average clock of all nodes, by exchanging local information between adjacent nodes using a *gossip* algorithm. This algorithm is robust to changes in the network topology, e.g., nodes entering or leaving the network.

Most of the above algorithms assume that the sensors do not move. Blind synchronization of moving sensors remains a challenging topic.

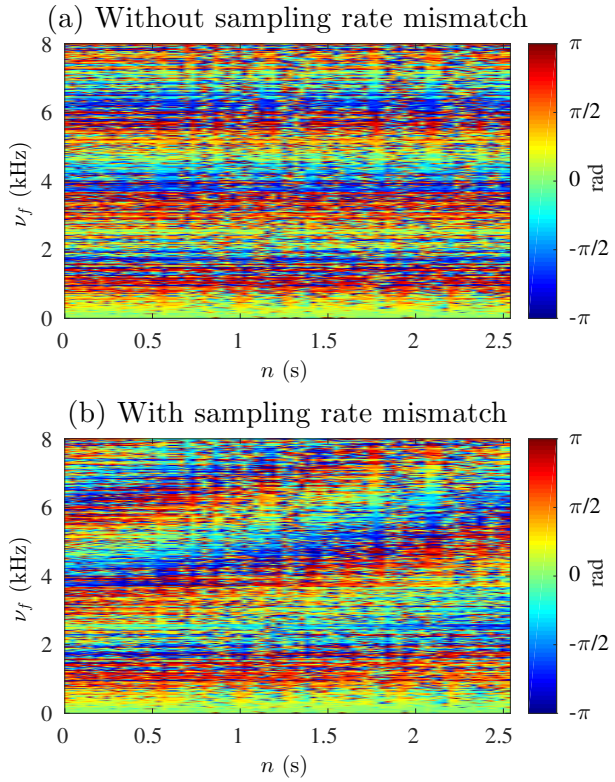

Figure 19.3 IPD between two microphones spaced by 15 cm belonging (a) to the same device or (b) to two distinct devices with $\epsilon = 6.25 \times 10^{-5}$ relative sampling rate mismatch. For illustration purposes, the recorded sound scene consists of a single speech source at a distance of 1.70 m and a DOA of 10° in a room with a reverberation time of 230 ms, without any interference or noise, and the two devices have zero temporal offset at $n = 0$.

19.4.2 Distributed algorithms

A second challenge raised by wireless acoustic sensor networks is that processing must be performed at each node without requiring the transmission of all signals to a master node. Markovich-Golan *et al.* (2015) reviewed three families of algorithms for distributed MVDR and linearly constrained minimum variance (LCMV) beamforming that rely on transmitting a compressed or fused version of the signals between neighboring nodes. One such family called *distributed adaptive node-specific signal estimation* (Bertrand and Moonen, 2010) also allows for distributed implementation of the MWF (Doclo *et al.*, 2009). Several network topologies can be handled, e.g., fully-connected or tree-structured, and the algorithms are shown to deliver equivalent results to centralized processing where a single processor has access to all signals. Efficient adaptation mechanisms can be designed to adapt to changes in the number of

available nodes and signals of interest (Markovich-Golan *et al.*, 2012b). Distributed implementations of MVDR and DS beamforming based on message passing, diffusion adaptation or randomized gossip algorithms have also been introduced (Heusdens *et al.*, 2012; O'Connor and Kleijn, 2014; Zeng and Hendriks, 2014). Gaubitch *et al.* (2014) describe a practical setup with smartphones.

Aside these advances on distributed beamforming, the distributed implementation of other families of speech enhancement and source separation methods remains an open problem. Souden *et al.* (2014) and Dorfan *et al.* (2015) made a step towards that goal by proposing distributed clustering schemes integrating intra- and internode location features for speech separation in wireless acoustic sensor networks.

19.4.3 Multimodal source separation and enhancement

Although we have focused on the audio modality, one must bear in mind that microphones are often embedded in devices equipped with other sensors. We have seen in Section 17.5 how video can be used to detect, localize, and extract relevant features for speech enhancement and separation. The application of these ideas to other sounds besides speech is a promising research direction: audiovisual processing is still in its infancy for music or environmental sound scenarios and it has been applied to related tasks only so far, e.g., (Joly *et al.*, 2016; Dinesh *et al.*, 2017). Multimodal separation and enhancement using other sensors, e.g., electroencephalogram (Das *et al.*, 2016), accelerometers (Zohourian and Martin, 2016), or laser rangefinders, remains largely untouched despite its great promise.

19.5 Towards widespread commercial use

Today, source separation and speech enhancement technology can be found in smartphones, hearing aids, and voice command systems. Their extension to new application scenarios raises several research issues.

19.5.1 Practical deployment constraints

Generally speaking, commercial source separation and speech enhancement systems are expected to incur a small memory footprint and a low computational cost. Noniterative learning-free algorithms such as various forms of beamforming score well according to these two criteria, hence their popularity in hearing aids (see Chapter 18). With the advent of more powerful, energy-efficient storage and processors, learning-based algorithms are becoming feasible in an increasing number of scenarios. For instance, Virtanen *et al.* (2013) evaluated the complexity of several NMF algorithms with various dictionary sizes. Efforts are also underway to reduce the complexity of DNNs, which are appealing compared to NMF due to the noniterative

nature of the forward pass. Kim and Smaragdis (2015) devised a bitwise neural network architecture, which showcases a comparable sound quality to a comprehensive real-valued DNN while spending significantly less memory and power.

In addition to the above constraints, several scenarios require processing the input signal in an online manner, that is without using future values of the mixture signal to estimate the source signals at a given time, and with low latency. Two strategies can be adopted to process the input signal as a stream: processing sliding blocks of data consisting of a few time frames (Mukai *et al.*, 2003; Joder *et al.*, 2012) or using exponential decay averaging (Gannot *et al.*, 1998; Rickard *et al.*, 2001; Lefèvre *et al.*, 2011; Schwartz *et al.*, 2015). Simon and Vincent (2012) combined both approaches into a single algorithm in the case of multichannel NMF. Whenever a very low latency is required, filtering the data in the time domain helps (Sunohara *et al.*, 2017).

19.5.2 Quality assessment

In order to assess the suitability of a separation or enhancement algorithm for a given application scenario and to keep improving it by extensive testing, accurate quality metrics are required. The metrics used today, as reviewed in Section 1.2.6, correlate with sound quality and speech intelligibility only to a limited extent. The composite metric of Loizou (2007) and the percetually-motivated metrics of Emiya *et al.* (2011) improve the correlation by training a linear regressor or a neural network on subjective quality scores collected from human subjects, but they suffer from the limited amount of subjective scores available for training today. Crowdsourcing is a promising way of collecting subjective scores for more audio data and from more subjects (Cartwright *et al.*, 2016) and increasing the accuracy of such machine learning-based quality metrics. Another issue is that sound quality and speech intelligibility are perceived differently by hearing-impaired vs. normal-hearing individuals, but also by different normal-hearing individuals (Emiya *et al.*, 2011). User-dependent quality metrics are an interesting research direction. Finally, most sound quality metrics are intrusive, in the sense that they require the true target signal in addition to the estimated signal. Developing nonintrusive quality metrics is essential to assess performance in many real scenarios where the true target signal is not available.

19.5.3 New application areas

Besides the applications to speech and music reviewed in Chapters 16, 17, 18, source separation and speech enhancement are being applied to an increasing range of application scenarios, such as enhancing the intelligibility of spoken dialogues in television broadcast (Geiger *et al.*, 2015), reducing the ego-noise due to fans and actuators in assistive robots (Ince *et al.*, 2009), rendering real sound scenes recorded via a microphone array in 3D over headphones in the context of virtual reality or augmented reality (Nikunen *et al.*, 2016), and encoding a recording into parametric source signals (Liutkus *et al.*, 2013) or sound objects (Vincent and Plumbley, 2007) for audio

upmixing and remixing purposes. Some of these topics were studied many years ago for the first time but are still active today due to the lack of a fully satisfactory solution. Source separation is also useful every time one must analyze a sound scene consisting of multiple sources. Examples include recognizing overlapped environmental sound events (Heittola *et al.*, 2013), monitoring traffic (Toyoda *et al.*, 2016), and controlling the noise disturbance of wind turbines (Dumortier *et al.*, 2017). These and other emerging application areas will further increase the widespread commercial use of source separation and speech enhancement technology.

Acknowledgment

We thank T. Gerkmann and R. Badeau for their input about the second section of this chapter.

Bibliography

Affes, S. and Grenier, Y. (1997) A signal subspace tracking algorithm for microphone array processing of speech. *IEEE Transactions on Speech and Audio Processing*, **5** (5), 425–437.

Asaei, A., Golbabaee, M., Bourlard, H., and Cevher, V. (2014) Structured sparsity models for reverberant speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22** (3), 620–633.

Attias, H. (2003) New EM algorithms for source separation and deconvolution with a microphone array, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, vol. V, vol. V, pp. 297–300.

Avargel, Y. and Cohen, I. (2008) Adaptive system identification in the short-time Fourier transform domain using cross-multiplicative transfer function approximation. *IEEE Transactions on Audio, Speech, and Language Processing*, **16** (1), 162–173.

Badeau, R. (2011) Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF), in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 253–256.

Bertrand, A. and Moonen, M. (2010) Distributed adaptive node-specific signal estimation in fully connected sensor networks — part i: Sequential node updating. *IEEE Transactions on Signal Processing*, **58** (10), 5277–5291.

Bronson, J. and Depalle, P. (2014) Phase constrained complex NMF: Separating overlapping partials in mixtures of harmonic musical sources, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 7475–7479.

Cartwright, M., Pardo, B., Mysore, G.J., and Hoffman, M. (2016) Fast and easy crowdsourced perceptual audio evaluation, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 619–623.

Chazan, S.E., Goldberger, J., and Gannot, S. (2016) A hybrid approach for speech enhancement using MoG model and neural network phoneme classifier. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24** (12), 2516–2530.

Cherkassky, D. and Gannot, S. (2017) Blind synchronization in wireless acoustic sensor networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **25** (3), 651–661.

Das, N., Van Eyndhoven, S., Francart, T., and Bertrand, A. (2016) Adaptive attention-driven speech enhancement for EEG-informed hearing prostheses, in *Proceedings of Annual International Conference of the IEEE Engineering in*

Medicine and Biology Society, pp. 77–80.

Deleforge, A., Forbes, F., and Horaud, R. (2015) Acoustic space learning for sound-source separation and localization on binaural manifolds. *International Journal of Neural Systems*, **25** (1). 1440003.

Dinesh, K., Li, B., Liu, X., Duan, Z., and Sharma, G. (2017) Visually informed multi-pitch analysis of string ensembles, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*.

Doclo, S., Moonen, M., Van den Bogaert, T., and Wouters, J. (2009) Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids. *IEEE Transactions on Audio, Speech, and Language Processing*, **17** (1), 38–51.

Dorfan, Y., Cherkassky, D., and Gannot, S. (2015) Speaker localization and separation using distributed expectation-maximization, in *Proceedings of European Signal Processing Conference*, pp. 1256–1260.

Dumortier, B., Vincent, E., and Deaconu, M. (2017) Recursive Bayesian estimation of the acoustic noise emitted by wind farms, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*.

Duong, N.Q.K., Tachibana, H., Vincent, E., Ono, N., Gribonval, R., and Sagayama, S. (2011) Multichannel harmonic and percussive component separation by joint modeling of spatial and spectral continuity, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 205–208.

Emiya, V., Vincent, E., Harlander, N., and Hohmann, V. (2011) Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, **19** (7), 2046–2057.

Fischer, D. and Gerkmann, T. (2016) Single-microphone speech enhancement using MVDR filtering and Wiener post-filtering, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 201–205.

Gannot, S., Burshtein, D., and Weinstein, E. (1998) Iterative and sequential Kalman filter-based speech enhancement algorithms. *IEEE Transactions on Speech and Audio Processing*, **6** (4), 373–385.

Gaubitch, N.D., Martinez, J., Kleijn, W.B., and Heusdens, R. (2014) On near-field beamforming with smartphone-based ad-hoc microphone arrays, in *Proceedings of International Workshop on Acoustic Echo and Noise Control*, pp. 94–98.

Geiger, J.T., Grosche, P., and Lacouture Parodi, Y. (2015) Dialogue enhancement of stereo sound, in *Proceedings of European Signal Processing Conference*, pp. 874–878.

Gerkmann, T. (2014) Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase. *IEEE Transactions on Signal Processing*, **62** (16), 4199–4208.

Gerkmann, T., Krawczyk-Becker, M., and Le Roux, J. (2015) Phase processing for single-channel speech enhancement: History and recent advances. *IEEE Signal Processing Magazine*, **32** (2), 55–66.

Griffin, D.W. and Lim, J.S. (1984) Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **32** (2), 236–243.

Gunawan, D. and Sen, D. (2010) Iterative phase estimation for the synthesis of separated sources from single-channel mixtures. *IEEE Signal Processing Letters*, **17** (5), 421–424.

Heittola, T., Mesaros, A., Virtanen, T., and Gabbouj, M. (2013) Supervised model training for overlapping sound events based on unsupervised source separation, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 8677–8681.

Hershey, J.R., Chen, Z., Le Roux, J., and Watanabe, S. (2016) Deep clustering: Discriminative embeddings for segmentation and separation, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 31–35.

Heusdens, R., Zhang, G., Hendriks, R.C., Zeng, Y., and Kleijn, W.B. (2012) Distributed MVDR beamforming for (wireless) microphone networks using message passing, in *Proceedings of International Workshop on Acoustic Echo and Noise Control*, pp. 1–4.

Heymann, J., Drude, L., and Haeb-Umbach, R. (2016) Neural network based spectral mask estimation for acoustic beamforming, in

Proceedings of IEEE International Conference on Audio, Speech and Signal Processing, pp. 196–200.

Higuchi, T., Takamune, N., Nakamura, T., and Kameoka, H. (2014) Underdetermined blind separation and tracking of moving sources based on DOA-HMM, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 3215–3219.

Hu, K. and Wang, D.L. (2013) An unsupervised approach to cochannel speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, **21** (1), 122–131.

Huang, H., Zhao, L., Chen, J., and Benesty, J. (2014) A minimum variance distortionless response filter based on the bifrequency spectrum for single-channel noise reduction. *Digital Signal Processing*, **33**, 169–179.

Huang, Y. and Benesty, J. (2012) A multi-frame approach to the frequency-domain single-channel noise reduction problem. *IEEE Transactions on Audio, Speech, and Language Processing*, **20** (4), 1256–1269.

Ince, G., Nakadai, K., Rodemann, T., Hasegawa, Y., Tsujino, H., and Imura, J. (2009) Ego noise suppression of a robot using template subtraction, in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 199–204.

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

Jaureguiberry, X., Vincent, E., and Richard, G. (2016) Fusion methods for speech enhancement and audio source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24** (7), 1266–1279.

Joder, C., Weninger, F., Eyben, F., Virette, D., and Schuller, B. (2012) Real-time speech separation by semi-supervised nonnegative matrix factorization, in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation*, pp. 322–329.

Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Champ, J., Planqué, R., Palazzo, S., and Müller, H. (2016) LifeCLEF 2016: Multimedia life species identification challenges, in *Proceedings of International Conference of the CLEF Association*, pp. 286–310.

Kameoka, H., Ono, N., Kashino, K., and Sagayama, S. (2009) Complex NMF: A new sparse representation for acoustic signals, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 3437–3440.

Kameoka, H., Yoshioka, T., Hamamura, M., Le Roux, J., and Kashino, K. (2010) Statistical model of speech signals based on composite autoregressive system with application to blind source separation, in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation*, pp. 245–253.

Kato, A. and Milner, B. (2016) HMM-based speech enhancement using sub-word models and noise adaptation, in *Proceedings of Interspeech*, pp. 3748–3752.

Kim, M. and Smaragdis, P. (2015) Bitwise neural networks, in *Proceedings of International Conference on Machine Learning Workshop on Resource-Efficient Machine Learning*.

Kolbæk, M., Tan, Z.H., and Jensen, J. (2017) Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **25** (1), 149–163.

Koldovský, Z., Málek, J., Tichavský, P., and Nesta, F. (2013) Semi-blind noise extraction using partially known position of the target source. *IEEE Transactions on Audio, Speech, and Language Processing*, **21** (10), 2029–2041.

Kounades-Bastian, D., Girin, L., Alameda-Pineda, X., Gannot, S., and Horaud, R. (2016) A variational EM algorithm for the separation of time-varying convolutive audio mixtures. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24** (8), 1408–1423.

Krawczyk, M. and Gerkmann, T. (2014) STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22** (12), 1931–1940.

Laufer-Goldshtein, B., Talmon, R., and Gannot, S. (2016) Semi-supervised sound source localization based on manifold

regularization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24** (8), 1393–1407.

Le, T.K. and Ono, N. (2017) Closed-form and near closed-form solutions for TDOA-based joint source and sensor localization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **65** (5), 1207–1221.

Le Roux, J., Kameoka, H., Ono, N., and Sagayama, S. (2010) Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency, in *Proceedings of International Conference on Digital Audio Effects*, pp. 1–7.

Le Roux, J. and Vincent, E. (2013) Consistent Wiener filtering for audio source separation. *IEEE Signal Processing Letters*, **20** (3), 217–220.

Lefèvre, A., Bach, F., and Févotte, C. (2011) Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence, in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 313–316.

Li, B., Sainath, T.N., Weiss, R.J., Wilson, K.W., and Bacchiani, M. (2016) Neural network adaptive beamforming for robust multichannel speech recognition, in *Proceedings of Interspeech*, pp. 1976–1979.

Liutkus, A., Durrieu, J.L., Daudet, L., and Richard, G. (2013) An overview of informed audio source separation, in *Proceedings of International Workshop on Image Analysis for Multimedia Interactive Services*, pp. 1–4.

Loizou, P.C. (2007) *Speech Enhancement: Theory and Practice*, CRC Press.

Lösch, B. and Yang, B. (2009) Online blind source separation based on time-frequency sparseness, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 117–120.

Madhu, N. and Martin, R. (2011) A versatile framework for speaker separation using a model-based speaker localization approach. *IEEE Transactions on Audio, Speech, and Language Processing*, **19** (7), 1900–1912.

Magron, P., Badeau, R., and David, B. (2015) Phase reconstruction of spectrograms based on a model of repeated audio events, in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–5.

Markovich-Golan, S., Bertrand, A., Moonen, M., and Gannot, S. (2015) Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks. *Signal Processing*, **107**, 4–20.

Markovich-Golan, S., Gannot, S., and Cohen, I. (2010) Subspace tracking of multiple sources and its application to speakers extraction, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 201–204.

Markovich-Golan, S., Gannot, S., and Cohen, I. (2012a) Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming, in *Proceedings of International Workshop on Acoustic Echo and Noise Control*.

Markovich-Golan, S., Gannot, S., and Cohen, I. (2012b) Low-complexity addition or removal of sensors/constraints in LCMV beamformers. *IEEE Transactions on Signal Processing*, **60** (3), 1205–1214.

Mignot, R., Chardon, G., and Daudet, L. (2014) Low frequency interpolation of room impulse responses using compressed sensing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22** (1), 205–216.

Miyabe, S., Ono, N., and Makino, S. (2015) Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation. *Signal Processing*, **107**, 185–196.

Mowlaee, P. and Saeidi, R. (2013) Iterative closed-loop phase-aware single-channel speech enhancement. *IEEE Signal Processing Letters*, **20** (12), 1235–1239.

Mowlaee, P., Saeidi, R., and Stylianou, Y. (2016) Advances in phase-aware signal processing in speech communication. *Speech Communication*, **81**, 1–29.

Mukai, R., Sawada, H., Araki, S., and Makino, S. (2003) Robust real-time blind source separation for moving speakers in a room, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. V–469–472.

Nickel, R.M., Astudillo, R.F., Kolossa, D., and

Martin, R. (2013) Corpus-based speech enhancement with uncertainty modeling and cepstral smoothing. *IEEE Transactions on Audio, Speech, and Language Processing*, **21** (5), 983–997.

Nikunen, J., Diment, A., Virtanen, T., and Vilermo, M. (2016) Binaural rendering of microphone array captures based on source separation. *Speech Communication*, **76**, 157–169.

Nugraha, A.A., Liutkus, A., and Vincent, E. (2016a) Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24** (10), 1652–1664.

Nugraha, A.A., Liutkus, A., and Vincent, E. (2016b) Multichannel music separation with deep neural networks, in *Proceedings of European Signal Processing Conference*, pp. 1748–1752.

O'Connor, M. and Kleijn, W.B. (2014) Diffusion-based distributed MVDR beamformer, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 810–814.

Pertilä, P. (2013) Online blind speech separation using multiple acoustic speaker tracking and time-frequency masking. *Computer Speech and Language*, **27** (3), 683–702.

Pertilä, P., Hämäläinen, M.S., and Mieskolainen, M. (2013) Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **21** (11), 2393–2402.

Rickard, S.J., Balan, R.V., and Rosca, J.P. (2001) Real-time time-frequency based blind source separation, in *Proceedings of International Conference on Independent Component Analysis and Signal Separation*, pp. 421–426.

Schasse, A. and Martin, R. (2014) Estimation of subband speech correlations for noise reduction via MVDR processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22** (9), 1355–1365.

Schmalenstroeer, J., Jebramcik, P., and Haeb-Umbach, R. (2015) A combined hardware–software approach for acoustic sensor network synchronization. *Signal Processing*, **107**, 171–184.

Schwartz, B., Gannot, S., and Habets, E.A.P. (2015) On-line speech dereverberation using Kalman filter and EM algorithm. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 394–406.

Shivakumar, P.G. and Georgiou, P. (2016) Perception optimized deep denoising autoencoders for speech enhancement, in *Proceedings of Interspeech*, pp. 3743–3747.

Simon, L.S.R. and Vincent, E. (2012) A general framework for online audio source separation, in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation*, pp. 397–404.

Sivasankaran, S., Vincent, E., and Illina, I. (2017) Discriminative importance weighting of augmented training data for acoustic model training, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*.

Souden, M., Kinoshita, K., Delcroix, M., and Nakatani, T. (2014) Location feature integration for clustering-based speech separation in distributed microphone arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22** (2), 354–367.

Stark, A.P. and Paliwal, K.K. (2008) Speech analysis using instantaneous frequency deviation, in *Proceedings of Interspeech*, pp. 2602–2605.

Sturmel, N. and Daudet, L. (2012) Iterative phase reconstruction of Wiener filtered signals, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 101–104.

Sunohara, M., Haruta, C., and Ono, N. (2017) Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of non-causal components, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*.

Talmon, R., Cohen, I., and Gannot, S. (2009) Convolutive transfer function generalized sidelobe canceler. *IEEE Transactions on Audio, Speech, and Language Processing*, **17** (7), 1420–1434.

Talmon, R. and Gannot, S. (2013) Relative transfer function identification on manifolds for supervised GSC beamformers, in *Proceedings of European Signal Processing*

Conference, pp. 1–5.

Thiergart, O., Taseska, M., and Habets, E.A.P. (2014) An informed parametric spatial filter based on instantaneous direction-of-arrival estimates. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22** (12), 2182–2196.

Toyoda, T., Ono, N., Miyabe, S., Yamada, T., and Makino, S. (2016) Vehicle counting and lane estimation with ad-hoc microphone array in real road environments, in *Proceedings of International Workshop on Nonlinear Circuits, Communications and Signal Processing*, pp. 622–625.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016) Wavenet: A generative model for raw audio, arXiv:1609.03499.

Vincent, E. and Plumbley, M.D. (2007) Low bit-rate object coding of musical audio using Bayesian harmonic models. *IEEE Transactions on Audio, Speech, and Language Processing*, **15** (4), 1273–1282.

Vincent, E., Watanabe, S., Nugraha, A.A., Barker, J., and Marxer, R. (2017) An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech and Language*, **46**, 535–557.

Virtanen, T., Gemmeke, J.F., and Raj, B. (2013) Active-set Newton algorithm for overcomplete non-negative representations of audio. *IEEE Transactions on Audio, Speech, and Language Processing*, **21** (11), 2277–2289.

Wang, L. and Doclo, S. (2016) Correlation maximization based sampling rate offset estimation for distributed microphone arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24** (3), 571–582.

Wang, Z., Vincent, E., and Serizel, R. (in preparation) Relative transfer function inverse regression from low dimensional manifolds. *IEEE Signal Processing Letters*.

Wehr, S., Lombard, A., Buchner, H., and Kellermann, W. (2007) "Shadow BSS" for blind source separation in rapidly time-varying acoustic scenes, in *Proceedings of International Conference on Independent Component Analysis and Signal Separation*, pp. 560–568.

Xiao, X., Xu, C., Zhang, Z., Zhao, S., Sun, S., Watanabe, S., Wang, L., Xie, L., Jones, D.L., Chng, E.S., and Li, H. (2016) A study of learning based beamforming methods for speech recognition, in *Proceedings of International Workshop on Speech Processing in Everyday Environments*, pp. 26–31.

Yegnanarayana, B. and Murthy, H. (1992) Significance of group delay functions in spectrum estimation. *IEEE Transactions on Signal Processing*, **40** (9), 2281–2289.

Yu, D., Kolbæk, M., Tan, Z.H., and Jensen, J. (2016) Permutation invariant training of deep models for speaker-independent multi-talker speech separation, arXiv:1607.00325.

Zagoruyko, S. and Komodakis, N. (2016) Wide residual networks, arXiv:1605.07146.

Zeng, Y. and Hendriks, R. (2014) Distributed delay and sum beamformer for speech enhancement via randomized gossip. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22** (1), 260–273.

Zhang, X., Zhang, H., Nie, S., Gao, G., and Liu, W. (2016) A pairwise algorithm using the deep stacking network for speech separation and pitch estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24** (6), 1066–1078.

Zohourian, M. and Martin, R. (2016) Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 430–434.