Introduction

Emmanuel Vincent, Sharon Gannot, Tuomas Virtanen

1
Introduction

Emmanuel Vincent, Sharon Gannot, and Tuomas Virtanen

Source separation and speech enhancement are core problems in the field of audio signal processing, with applications to speech, music, and environmental audio. Research in this field has accompanied technological trends, such as the move from landline to mobile or hands-free phones, the gradual replacement of stereo by 3D audio, and the emergence of connected devices equipped with one or more microphones that can execute audio processing tasks which were previously regarded as impossible. In this short introductory chapter, after a brief discussion of the application needs in Section 1.1, we define the problems of source separation and speech enhancement and introduce relevant terminology regarding the scenarios and the desired outcome in Section 1.2. We then present the general processing scheme followed by most source separation and speech enhancement approaches and categorize these approaches in Section 1.3. Finally, we provide an outline of this book in Section 1.4.

1.1
Why are source separation and speech enhancement needed?

The problems of source separation and speech enhancement arise from several application needs in the context of speech, music, or environmental audio processing.

Real-world speech signals are often contaminated by interfering speakers, environmental noise, and/or reverberation. These phenomena deteriorate speech quality and, in adverse scenarios, speech intelligibility and automatic speech recognition (ASR) performance. Source separation and speech enhancement are therefore required in such scenarios. For instance, spoken communication over mobile phones or hands-free systems requires the separation or enhancement of the near-end speaker's voice with respect to interfering speakers and environmental noises before it is transmitted to the far-end listener. Conference call systems or hearing aids face the same problem, except that several speakers may be considered as targets. Source separation and speech enhancement are also crucial preprocessing steps for robust distant-microphone ASR, as available in today's personal assistants, car navigation systems,

televisions, video game consoles, medical dictation devices, and meeting transcription systems. Finally, they are necessary components in providing humanoid robots, assistive listening devices, and surveillance systems with "super-hearing" capabilities, which may exceed the hearing capabilities of humans.

Besides speech, music and movie soundtracks are another important application area for source separation. Indeed, music recordings typically involve several instruments playing together live or mixed together in a studio, while movie soundtracks involve speech overlapped with music and sound effects. Source separation has been successfully used to upmix mono or stereo recordings to 3D sound formats and/or to remix them. It lies at the core of object-based audio coders, which encode a given recording as the sum of several sound objects, that can then easily be rendered and manipulated. It is also useful for music information retrieval purposes, e.g., to transcribe the melody or the lyrics of a song from the separated singing voice.

An emerging research field with many real-life applications concerns the analysis of general sound scenes, involving the detection of sound events, their localization and tracking, and the inference of the acoustic environment properties.

1.2 What are the goals of source separation and speech enhancement?

The goal of source separation and speech enhancement can be defined in layman terms as that of recovering the signal of one or more sound sources from an observed signal involving other sound sources and/or reverberation. This definition turns out to be ambiguous. In order to address the ambiguity, the notion of source and the process leading to the observed signal must be characterized more precisely. In this section and in the rest of this book, we adopt the general notations defined on p. xv–xvii.

1.2.1 Single-channel vs. multichannel

Let us assume that the observed signal has I *channels* indexed by $i \in \{1, \ldots, I\}$. By channel, we mean the output of one microphone in the case when the observed signal has been recorded by one or more microphones, or the input of one loudspeaker in the case when it is destined to be played back on one or more loudspeakers[1)]. A signal with $I = 1$ channel is called *single-channel* and represented by a scalar $x(t)$, while a signal with $I > 1$ channels is called *multichannel* and represented by an $I \times 1$ vector $\mathbf{x}(t)$. The explanation below employs multichannel notation, but is also valid in the single-channel case.

1) This is the usual meaning of "channel" in the field of professional and consumer audio. In the field of telecommunications and, by extension, in some speech enhancement papers, "channel" refers to the distortions (e.g., noise and reverberation) occurring when transmitting a signal instead. The latter meaning will not be employed hereafter.

1.2.2 Point vs. diffuse sources

Furthermore, let us assume that there are J sound *sources* indexed by $j \in \{1, \ldots, J\}$. The word "source" can refer to two different concepts. A *point source* such as a human speaker, a bird, or a loudspeaker is considered to emit sound from a single point in space. It can be represented as a single-channel signal. A *diffuse source* such as a car, a piano, or rain simultaneously emits sound from a whole region in space. The sounds emitted from different points of that region are different but not always independent of each other. Therefore, a diffuse source can be thought of as an infinite collection of point sources. The estimation of the individual point sources in this collection can be important for the study of vibrating bodies, but it is considered irrelevant for source separation or speech enhancement. A diffuse source is therefore typically represented by the corresponding signal recorded at the microphone(s) and it is processed as a whole.

1.2.3 Mixing process

The mixing process leading to the observed signal can generally be expressed in two steps. First, each single-channel point source signal $s_j(t)$ is transformed into an $I \times 1$ source *spatial image* signal (Vincent *et al.*, 2012) $\mathbf{c}_j(t) = [c_{1j}(t), \ldots, c_{Ij}(t)]^T$ by means of a possibly nonlinear spatialization operation. This operation can describe the acoustic propagation from the point source to the microphone(s), including reverberation, or some artificial mixing effects. Diffuse sources are directly represented by their $I \times 1$ spatial images $\mathbf{c}_j(t)$ instead. Second, the spatial images of all sources are summed to yield the observed $I \times 1$ signal $\mathbf{x}(t) = [x_1(t), \ldots, x_I(t)]^T$ called *mixture*:

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t). \tag{1.1}$$

This summation is due to the superposition of the sources in the case of microphone recording or to explicit summation in the case of artificial mixing. This implies that the spatial image of each source represents the contribution of the source to the mixture signal. A schematic overview of the mixing process is depicted in Fig. 1.1. More specific details are given in Chapter 3.

Note that *target* sources, *interfering* sources, and *noise* are treated in the same way in this formulation. All these signals can be either point or diffuse sources. The choice of target sources depends on the use case. Also, the distinction between interfering sources and noise may or may not be relevant depending on the use case. In the context of speech processing, these terms typically refer to undesired speech vs. nonspeech sources, respectively. In the context of music or environmental sound processing, this distinction is most often irrelevant and the former term is preferred to the latter.

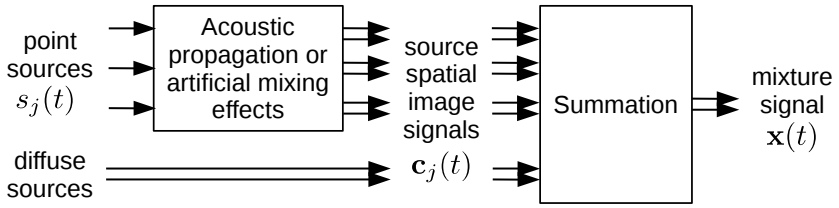

Figure 1.1 General mixing process, illustrated in the case of $J = 4$ sources, including 3 point sources and 1 diffuse source, and $I = 2$ channels.

In the following, we assume that all signals are digital, meaning that the time variable t is discrete. We also assume that quantization effects are negligible, so that we can operate on continuous amplitudes. Regarding the conversion of acoustic signals to analog audio signals and analog signals to digital, see, e.g., Havelock *et al.* (2008, Part XII) and Pohlmann (1995, pp. 22–49).

1.2.4 Separation vs. enhancement

The above mixing process implies one or more distortions of the target signals: interfering sources, noise, reverberation, and echo emitted by the loudspeakers (if any). In this context, *source separation* refers to the problem of extracting one or more target sources while suppressing interfering sources and noise. It explicitly excludes dereverberation and echo cancellation. *Enhancement* is more general, in that it refers to the problem of extracting one or more target sources while suppressing all types of distortion, including reverberation and echo. In practice, though, this term is mostly used in the case when the target sources are speech. In the audio processing literature, these two terms are often interchanged, especially when referring to the problem of suppressing both interfering speakers and noise from a speech signal. Note that, for either source separation or enhancement tasks, the extracted source(s) can be either the spatial image of the source or its direct path component, namely the delayed and attenuated version of the original source signal (Vincent *et al.*, 2012; Gannot *et al.*, 2001).

The problem of echo cancellation is out of the scope of this book. Please refer to Hänsler and Schmidt (2004) for a comprehensive overview of this topic. The problem of source localization and tracking cannot be viewed as a separation or enhancement task, but it is sometimes used as a preprocessing step prior to separation or enhancement, hence it is discussed in Chapter 4. Dereverberation is explored in Chapter 15. The remaining chapters focus on separation and enhancement.

1.2.5 Typology of scenarios

The general source separation literature has come up with a terminology to characterize the mixing process (Hyvärinen *et al.*, 2001; O'Grady *et al.*, 2005; Comon and Jutten, 2010). A given mixture signal is said to be

- *linear* if the mixing process is linear, and *nonlinear* otherwise,
- *time-invariant* if the mixing process is fixed over time, and *time-varying* otherwise,
- *instantaneous* if the mixing process simply scales each source signal by a different factor on each channel, *anechoic* if it also applies a different delay to each source on each channel, and *convolutive* in the more general case when it results from summing multiple scaled and delayed versions of the sources,
- *overdetermined* if there is no diffuse source and the number of point sources is strictly smaller than the number of channels, *determined* if there is no diffuse source and the number of point sources is equal to the number of channels, and *underdetermined* otherwise.

This categorization is relevant but has limited usefulness in the case of audio. As we shall see in Chapter 3, virtually all audio mixtures are linear (or can be considered so) and convolutive. The over- vs. underdetermined distinction was motivated by the fact that a determined or overdetermined linear time-invariant mixture can be perfectly separated by inverting the mixing system using a linear time-invariant inverse (see Chapter 13). In practice, however, the majority of audio mixtures involve at least one diffuse source (e.g., background noise) or more point sources than channels. Audio source separation and speech enhancement systems are therefore generally faced with underdetermined linear (time-invariant or time-varying) convolutive mixtures[2)].

Recently, an alternative categorization has been proposed based on the amount of prior information available about the mixture signal to be processed (Vincent *et al.*, 2014). The separation problem is said to be

- *blind* when absolutely no information is given about the source signals, the mixing process or the intended application,
- *weakly guided* or *semi-blind* when general information is available about the context of use, e.g., the nature of the sources (speech, music, environmental sounds), the microphone positions, the recording scenario (domestic, outdoor, professional music…), and the intended application (hearing aid, speech recognition…),
- *strongly guided* when specific information is available about the signal to be processed, e.g., the spatial location of the sources, their activity pattern, the identity

2) Certain authors call mixtures for which the number of point sources is equal to (resp. strictly smaller than) the number of channels as determined (resp. overdetermined) even when there is a diffuse noise source. Perfect separation of such mixtures cannot be achieved using time-invariant filtering anymore: it requires a time-varying separation filter, similarly to underdetermined mixtures. Indeed, a time-invariant filter can cancel the interfering sources and reduce the noise, but it cannot cancel the noise perfectly. We prefer the above definition of "determined" and "overdetermined", which matches the mathematical definition of these concepts for systems of linear equations and has a more direct implication on the separation performance achievable by linear time-invariant filtering.

of the speakers, or a musical score,
- *informed* when highly precise information about the sources and the mixing process is encoded and transmitted along with the audio.

Although the term "blind" has been extensively used in source separation (see Chapters 4, 10, 11, and 13), strictly blind separation is inapplicable in the context of audio. As we shall see in Chapter 13, certain assumptions about the probability distribution of the sources and/or the mixing process must always be made in practice. Strictly speaking, the term "weakly guided" would therefore be more appropriate. Informed separation is closer to audio coding than to separation and will be briefly covered in Chapter 16. All other source separation and speech enhancement methods reviewed in this book are therefore either weakly or strongly guided.

Finally, the separation or enhancement problem can be categorized depending on the order in which the samples of the mixture signal are processed. It is called *online* when the mixture signal is captured in real time by small blocks of a few tens or hundred samples and each block must be processed given past blocks only, or few future blocks introducing tolerated latency. On the contrary, it is called *offline* or *batch* when the recording has been completed and it is processed as a whole, using both past and future samples to estimate a given sample of the sources.

1.2.6 Evaluation

Using current technology, source separation and dereverberation are rarely perfect in real-life scenarios. For each source, the estimated source or source spatial image signal can differ from the true target signal in several ways, including (Vincent *et al.*, 2006; Loizou, 2007)

- *distortion* of the target signal, e.g., lowpass filtering, fluctuating intensity over time,
- residual interference or noise from the other sources,
- "*musical noise*" *artifacts*, i.e., isolated sounds in both frequency and time similar to those generated by a lossy audio codec at a very low bitrate.

The assessment of these distortions is essential to compare the merits of different algorithms and understand how to improve their performance.

Ideally, this assessment should be based on the performance of the tested source separation or speech enhancement method for the desired application. Indeed, the importance of various types of distortion depends on the specific application. For instance, some amount of distortion of the target signal which is deemed acceptable when listening to the separated signals can lead to a major drop in the speech recognition performance. Artifacts are often greatly reduced when the separated signals are remixed together in a different way, while they must be avoided at all costs in hearing aids. Standard performance metrics are typically available for each task, some of which will be mentioned later in this book.

When the desired application involves listening to the separated or enhanced signals or to a remix, sound quality and, whenever relevant, speech intelligibility should

Table 1.1 Evaluation software and metrics.

Software	Implemented metrics
ITU-T (2001)	PESQ
Taal *et al.* (2011)[3)]	STOI
Loizou (2007)[4)]	segmental SNR log-likelihood ratio cepstrum distance
BSS Eval (Vincent *et al.*, 2006)[5)]	SDR SIR SAR
Falk *et al.* (2010)	speech to reverberation modulation energy ratio

ideally be assessed by means of a subjective listening test (ITU-T, 2003; Emiya *et al.*, 2011; ITU-T, 2016). Contrary to a widespread belief, a number of subjects as low as ten can sometimes suffice to obtain statistically significant results. However, data selection and subject screening are time-consuming. Recent attempts with crowdsourcing are a promising way of making subjective testing more convenient in the near future (Cartwright *et al.*, 2016). An alternative approach is to use objective separation or dereverberation metrics. Table 1.1 provides an overview of some commonly used metrics. The so-called *PESQ* metric, the segmental signal-to-noise ratio (SNR), and the *signal-to-distortion ratio* (SDR) measure the overall estimation error, including the three types of distortion listed above. The so-called *STOI* index is more related to speech intelligibility by humans, and the log-likelihood ratio and cepstrum distance to ASR by machines. The *signal-to-interference ratio* (SIR) and the *signal-to-artifacts ratio* (SAR) aim to assess separately the latter two types of distortion listed above. The segmental SNR, SDR, SIR, and SAR are expressed in decibels (dB), while PESQ and STOI are expressed on a perceptual scale. More specific metrics will be reviewed later in the book.

A natural question that arises once the metrics have been defined is: what is the best performance possibly achievable for a given mixture signal? This can be used to assess the difficulty of solving the source separation or speech enhancement problem in a given scenario and the room left for performance improvement as compared to current systems. This question can be answered using *oracle* or *ideal* estimators based on the knowledge of the true source or source spatial image signals (Vincent *et al.*, 2007).

3) http://amtoolbox.sourceforge.net/doc/speech/taal2011.php
4) http://www.crcpress.com/product/isbn/9781466504219
5) http://bass-db.gforge.inria.fr/bss_eval/

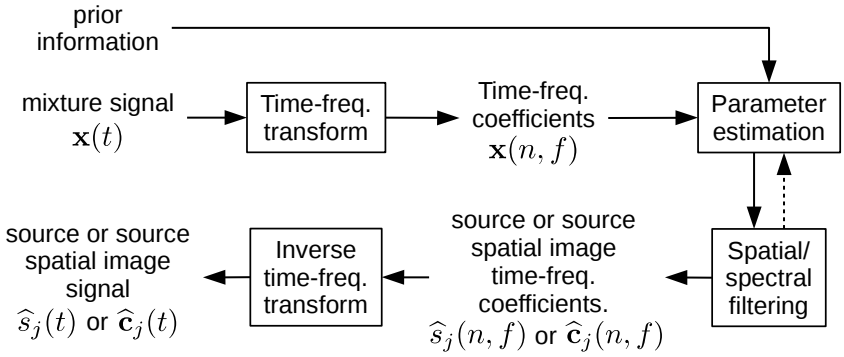

Figure 1.2 General processing scheme for single-channel and multichannel source separation and speech enhancement.

1.3 How can source separation and speech enhancement be addressed?

Now that we have defined the goals of source separation and speech enhancement, let us turn to how they can be addressed.

1.3.1 General processing scheme

Many different approaches to source separation and speech enhancement have been proposed in the literature. The vast majority of approaches follow the general processing scheme depicted in Fig. 1.2, which applies both to single-channel and multichannel scenarios. The time-domain mixture signal $\mathbf{x}(t)$ is represented in the time-frequency domain (see Chapter 2). A model of the complex-valued time-frequency coefficients of the mixture $\mathbf{x}(n,f)$ and the sources $s_j(n,f)$ (resp. the source spatial images $\mathbf{c}_j(n,f)$) is built. The choice of model is motivated by the general prior information about the scenario (see Section 1.2.5). The model parameters are estimated from $\mathbf{x}(n,f)$ or from separate training data according to a certain criterion. Additional specific prior information can be used to help parameter estimation whenever available. Given these parameters, a time-varying single-output (resp. multiple-output) complex-valued filter is derived and applied to the mixture $\mathbf{x}(n,f)$ in order to obtain an estimate of the complex-valued time-frequency coefficients of the sources $\widehat{s}_j(n,f)$ (resp. the source spatial images $\widehat{\mathbf{c}}_j(n,f)$). Finally, the time-frequency transform is inverted, yielding time-domain source estimates $\widehat{s}_j(t)$ (resp. source spatial image estimates $\widehat{\mathbf{c}}_j(t)$).

1.3.2 Converging historical trends

The various approaches proposed in the literature differ by the choice of model, the parameter estimation algorithm, and the derivation of the separation or enhancement filter. Research has followed three historical paths. First, microphone array processing emerged from the theory of sensor array processing for telecommunications and it focused mostly on the localization and enhancement of speech in noisy or reverberant environments. Second, the concepts of independent component analysis (ICA) and nonnegative matrix factorization (NMF) gave birth to a stream of blind source separation (BSS) methods aiming to address "cocktail party" scenarios (as coined by Cherry (1953)) involving several sound sources mixed together. Third, attempts to implement the sound segregation properties of the human ear (Bregman, 1994) in a computer gave rise to computational auditory scene analysis (CASA) methods. These paths have converged in the last decade and they are hardly distinguishable anymore. As a matter of fact, virtually all source separation and speech enhancement methods rely on modeling the *spectral* properties of the sources, i.e., their distribution of energy over time and frequency, and/or their *spatial* properties, i.e., the relations between channels over time.

Most books and surveys about audio source separation and speech enhancement so far have focused on a single point of view, namely microphone array processing (Gay and Benesty, 2000; Brandstein and Ward, 2001; Loizou, 2007; Cohen *et al.*, 2010), CASA (Divenyi, 2004; Wang and Brown, 2006), BSS (O'Grady *et al.*, 2005; Makino *et al.*, 2007; Virtanen *et al.*, 2015), or machine learning (Vincent *et al.*, 2010, 2014; Le Roux *et al.*, in preparation). These are complemented by books on general sensor array processing and BSS (Hyvärinen *et al.*, 2001; Van Trees, 2002; Cichocki *et al.*, 2009; Haykin and Liu, 2010; Comon and Jutten, 2010) which do not specifically focus on speech and audio, and books on general speech processing (Benesty *et al.*, 2007; Wölfel and McDonough, 2009; Virtanen *et al.*, 2012; Li *et al.*, 2015) which do not specifically focus on separation and enhancement. Few books and surveys have attempted to cross the boundaries between these points of view (Benesty *et al.*, 2005; Cohen *et al.*, 2009; Gannot *et al.*, 2017; Makino, in preparation), but they do not cover all state-of-the-art approaches and all application scenarios. We designed this book to provide the most comprehensive, up-to-date overview of the state-of-the-art and allow readers to acquire a wide understanding of these topics.

1.3.3 Typology of approaches

With the merging of the three historical paths introduced above, a new categorization of source separation and speech enhancement methods has become necessary. One of the most relevant ones today is based on the use of training data to estimate the model parameters and on the nature of this data. This categorization differs from the one in Section 1.2.5: it does not relate to the problem posed, but to the way it is solved. Both categorizations are essentially orthogonal. Le Roux *et al.* (in preparation) distinguish

four categories of approaches:

- *learning-free* methods do not rely on any training data: all parameters are either fixed manually by the user or estimated from the test mixture $\mathbf{x}(n, f)$ (e.g., frequency-domain ICA in Section 13.2);
- *unsupervised source modeling* methods train a model for each source from unannotated isolated signals of that source type, i.e., without using any information about each training signal besides the source type (e.g., so-called "supervised NMF" in Section 8.1.3);
- *supervised source modeling* methods train a model for each source from annotated isolated signals of that source type, i.e., using additional information about each training signal (e.g., isolated notes annotated with pitch information in the case of music, see Section 16.2.2.1);
- *separation based training* methods (e.g., deep neural network (DNN) based methods in Section 7.3) train a separation mechanism or jointly train models for all sources from mixture signals given the underlying true source signals.

In all cases, development data whose conditions are similar to the test mixture can be used to tune a small number of hyperparameters. Certain methods borrow ideas from several categories of approaches. For instance, "semi-supervised" NMF in Section 8.1.4 is halfway between learning-free and unsupervised source modeling based separation.

Other terms were used in the literature, such as generative vs. discriminative methods. We do not use these terms in the following and prefer the finer-grained categories above, which are specific to source separation and speech enhancement.

1.4 Outline

This book is structured in four parts.

Part I introduces the basic concepts of time-frequency processing in Chapter 2 and sound propagation in Chapter 3 and highlights the spectral and spatial properties of the sources. Chapter 4 provides additional background material on source activity detection and localization. These chapters are mostly designed for beginners and can be skipped by experienced readers.

Part II focuses on single-channel separation and enhancement based on the spectral properties of the sources. We first define the concept of spectral filtering in Chapter 5. We then explain how suitable spectral filters can be derived from various models and we present algorithms to estimate the model parameters in Chapters 6 to 9. Most of these algorithms are not restricted to a given application area.

Part III addresses multichannel separation and enhancement based on spatial and/or spectral properties. It follows a similar structure to Part II. We first define the concept of spatial filtering in Chapter 10 and proceed with several models and algorithms in Chapters 11 to 14. Chapter 15 focuses on dereverberation. Again, most of the algorithms reviewed in this part are not restricted to a given application area.

Readers interested in single-channel audio shall focus on Part II, while those interested in multichannel audio are advised to read both Parts II and III since most single-channel algorithms can be employed or extended in a multichannel context. In either case, Chapters 5 and 10 must be read first, since they are are prerequisites to the other chapters. Chapters 6 to 9 and 11 to 15 are independent of each other and can be read separately, except Chapter 9 which relies on Chapter 8. Reading all chapters in either part is strongly recommended, however. This will provide the reader with a more complete view of the field and allow him/her to select the most appropriate algorithm or develop a new algorithm for his own use case.

Part IV presents the challenges and opportunities associated with the use of these algorithms in specific application areas: music in Chapter 16, speech in Chapter 17, and hearing instruments in Chapter 18. These chapters are independent of each other and may be skipped or not depending on the reader's interest. We conclude by discussing several research perspectives in Chapter 19.

Bibliography

Benesty, J., Makino, S., and Chen, J. (eds) (2005) *Speech Enhancement*, Springer.

Benesty, J., Sondhi, M.M., and Huang, Y. (eds) (2007) *Springer Handbook of Speech Processing and Speech Communication*, Springer.

Brandstein, M.S. and Ward, D.B. (eds) (2001) *Microphone Arrays: Signal Processing Techniques and Applications*, Springer.

Bregman, A.S. (1994) *Auditory scene analysis: The perceptual organization of sound*, MIT Press.

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.

Cherry, E.C. (1953) Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, **25** (5), 975–979.

Cichocki, A., Zdunek, R., Phan, A.H., and Amari, S. (2009) *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley.

Cohen, I., Benesty, J., and Gannot, S. (2009) *Speech processing in modern communication: Challenges and perspectives*, vol. 3, Springer.

Cohen, I., Benesty, J., and Gannot, S. (eds) (2010) *Speech Processing in Modern Communication: Challenges and Perspectives*, Springer.

Comon, P. and Jutten, C. (eds) (2010) *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, Academic Press.

Divenyi, P. (ed.) (2004) *Speech Separation by Humans and Machines*, Springer.

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.

Falk, T.H., Zheng, C., and Chan, W.Y. (2010) A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, **18** (7), 1766–1774.

Gannot, S., Burshtein, D., and Weinstein, E. (2001) Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, **49** (8), 1614–1626.

Gannot, S., Vincent, E., Markovich-Golan, S., and Ozerov, A. (2017) A consolidated perspective on multi-microphone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **25** (4), 692–730.

Gay, S.L. and Benesty, J. (eds) (2000) *Acoustic Signal Processing for Telecommunication*, Kluwer.

Hänsler, E. and Schmidt, G. (2004) *Acoustic Echo and Noise Control: A Practical Approach*, Wiley.

Havelock, D., Kuwano, S., and Vorländer, M. (eds) (2008) *Handbook of Signal Processing in Acoustics*, vol. 2, Springer.

Haykin, S. and Liu, K.R. (eds) (2010) *Handbook on Array Processing and Sensor Networks*, Wiley.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001) *Independent Component Analysis*, Wiley.

ITU-T (2001) Recommendation P.862. perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.

ITU-T (2003) Recommendation P.835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm.

ITU-T (2016) Recommendation P.807. subjective test methodology for assessing speech intelligibility.

Le Roux, J., Erdogan, H., Hershey, J.R., Vincent, E., and Watanabe, S. (in preparation) Learning-based speech enhancement and separation: A state of the art. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Li, J., Deng, L., Haeb-Umbach, R., and Gong, Y. (2015) *Robust Automatic Speech Recognition*, Academic Press.

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

Makino, S. (ed.) (in preparation) *Audio Source Separation*, Springer.

Makino, S., Lee, T.W., and Sawada, H. (eds) (2007) *Blind Speech Separation*, Springer.

O'Grady, P.D., Pearlmutter, B.A., and Rickard, S.T. (2005) Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology*, **15**, 18–33.

Pohlmann, K.C. (1995) *Principles of Digital Audio*, McGraw-Hill, 3rd edn..

Taal, C.H., Hendriks, R.C., Heusdens, R., and Jensen, J. (2011) An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, **19** (7), 2125–2136.

Van Trees, H.L. (2002) *Optimum Array Processing*, Wiley.

Vincent, E., Araki, S., Theis, F.J., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, B.V., Lutter, D., and Duong, N.Q.K. (2012) The Signal Separation Evaluation Campaign (2007–2010): Achievements and remaining challenges. *Signal Processing*, **92**, 1928–1936.

Vincent, E., Bertin, N., Gribonval, R., and Bimbot, F. (2014) From blind to guided audio source separation: How models and side information can improve the separation of sound. *IEEE Signal Processing Magazine*, **31** (3), 107–115.

Vincent, E., Gribonval, R., and Févotte, C. (2006) Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, **14** (4), 1462–1469.

Vincent, E., Gribonval, R., and Plumbley, M.D. (2007) Oracle estimators for the benchmarking of source separation algorithms. *Signal Processing*, **87** (8), 1933–1950.

Vincent, E., Jafari, M.G., Abdallah, S.A., Plumbley, M.D., and Davies, M.E. (2010) Probabilistic modeling paradigms for audio source separation, in *Machine Audition: Principles, Algorithms and Systems*, IGI Global, pp. 162–185.

Virtanen, T., Gemmeke, J.F., Raj, B., and Smaragdis, P. (2015) Compositional models for audio processing: Uncovering the structure of sound mixtures. *IEEE Signal Processing Magazine*, **32** (2), 125–144.

Virtanen, T., Singh, R., and Raj, B. (eds) (2012) *Techniques for Noise Robustness in Automatic Speech Recognition*, Wiley.

Wang, D. and Brown, G.J. (eds) (2006) *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley.

Wölfel, M. and McDonough, J. (2009) *Distant Speech Recognition*, Wiley.