IMAGE LICENSED BY INGRAM PUBLISHING

Recent Developments in Speech Enhancement in the Short-Time Fourier Transform Domain

Mahdi Parchami, Wei-Ping Zhu, Benoit Champagne, and Eric Plourde

Abstract

In this paper, we present an overview on the topic of noise reduction in the short-time Fourier transform (STFT) domain. First, we briefly review the conventional literature in the single- and multi-channel cases separately. In the single-channel scenario, we focus on the spectral subtractive methods, Wiener filter based methods, speech amplitude estimators and estimators of the complex STFT coefficients. In the multi-channel scenario, we investigate in short a selection of key beamforming approaches as well as conventional post-filtering methods. Next, a detailed survey of the most recent advances in the STFT-based noise reduction methods is provided. This includes STSA estimators with super-Gaussian priors, noise power spectral density (PSD) estimation, estimation methods in the modulation domain, estimation of spectral phase and noise PSD matrix estimation for multi-channel applications. Finally, we summarize the presented material and draw important conclusions on each of the investigated topics.

Digital Object Identifier 10.1109/MCAS.2016.2583681
Date of publication: 19 August 2016

I. Introduction

The objective of speech enhancement is to improve the intelligibility and/or overall perceptual quality of degraded speech signals using signal processing techniques. Basically, the recorded speech signal in a real-world application may be corrupted by various noise types, interferences, echoes and reverberation resulting from the acoustic environment and enclosure. These degradations can significantly reduce the intelligibility of the speech signal by human listeners and also deteriorate the performance of speech coding and recognition systems [1], [2]. Hence, high performance speech enhancement techniques are necessary for all speech communication systems.

Fig. 1 illustrates a general scenario of capturing an audio signal where speech enhancement is required. As is observed, the captured signal can be generally corrupted

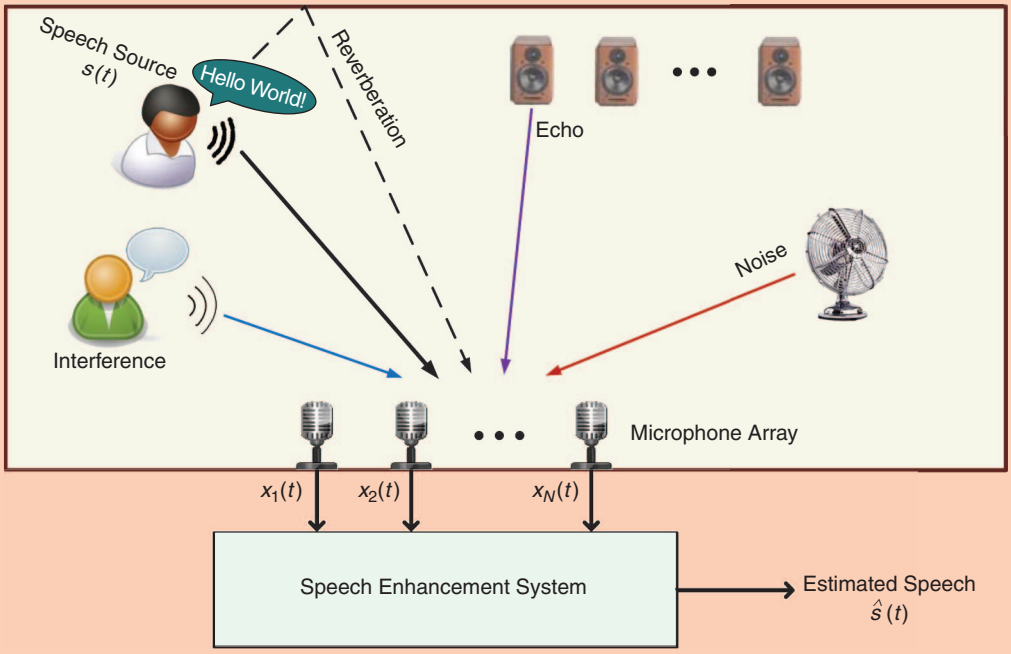

Figure 1. Microphone array speech enhancement in a general application.

by environmental noise, speech-like interferences, speech echoes and acoustical reverberation. These corruptions can distort the clean speech quality as well as its intelligibility, thus necessitating the implementation of a suitable speech enhancement algorithm on the corrupted speech. Enhancement of speech degraded by noise, or noise reduction, is the most important topic in speech enhancement. In general, noise reduction for speech signals is a difficult task to accomplish for many reasons. First, the nature and characteristics of the corrupting disturbances in speech can change dramatically in different environments or from one application to the other. Second, the performance criteria under which the fidelity of speech enhancement algorithms is evaluated can be different depending on the application. As a common example, in the single-channel (i.e. one-microphone) case where the speech degradation is due to uncorrelated additive noise, noise reduction can be generally achieved at the expense of introducing speech distortion. In this case, even though noise reduction measures demonstrate quality improvement in the processed speech, distortion measures for the latter can be worse than those of the noisy speech. In fact, there exists a compromise between the amounts of noise reduction achieved by conventional speech enhancement algorithms and the speech distortion introduced in the clean speech [3], [4].

The most important applications of noise reduction include mobile phones, voice over internet protocol (VoIP), teleconferencing systems, speech recognition, and hearing aids. Most voice processing and communication systems used in noisy environments highly require speech restoration modules in order to function properly. For instance, in digital telephony applications, ambient noise prevents the speech codecs from estimating the required spectral parameters accurately. Therefore, the resulting coded speech after transmission sounds distorted and still contains corrupting noise. Hence, to improve the performance of speech codecs, a speech enhancement subsystem has to be employed as a front-end to reduce the noise energy. Moreover, in automatic speech recognition (ASR), regardless of the performance of the underlying ASR system, the input speech quality hugely effects the speech recognition accuracy. Therefore, speech enhancement solutions play an important role in the overall performance of the ASR systems. Speech enhancement is also vital to hearing aid devices as these devices inherently amplify the present noise in the received audio, and therefore, may pose further difficulty in understanding

Mahdi Parchami and Wei-Ping Zhu are with the Department of Electrical and Computer Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal (H3G 1M8), Quebec, Canada (e-mails: {m_parch, weiping}@ece.concordia.ca). Benoit Champagne is with the Department of Electrical and Computer Engineering, McGill University, 3480 University Street, Montreal (H3A 0E9), Quebec, Canada (e-mail: benoit.champagne@mcgill.ca). Eric Plourde is with the Department of Electrical and Computer Engineering, Université de Sherbrooke, 2500 boul. de l'Université, Sherbrooke (J1K 2R1), Quebec, Canada (e-mail: eric.plourde@usherbrooke.ca).

voice to the hearing impaired [5], [6]. Thus, with the fast development of the aforementioned speech and audio systems, there will be a growing need for more efficient noise suppression algorithms in the future.

From a general point of view, the main noise reduction algorithms can be categorized into several fundamental classes including adaptive filtering methods, spectral subtractive algorithms, Wiener filtering and its variations, statistical model-based methods and subspace algorithms [7]. Whereas a performance comparison in terms of speech quality and intelligibility can be done amongst different categories of speech enhancement algorithms, factors such as computational load, need for training data and restrictive assumptions about noise and speech environments have to be taken into account in order to select the proper noise reduction method for a given application. While the noisy speech signal is usually available in the form of a sequence of time domain samples, speech enhancement often benefits from an implementation in a transform domain. The most important signal transformations in the field of speech processing include the discrete Fourier transform (DFT), discrete wavelet transform, discrete cosine transform (DCT) and Karhunen-Loeve transform (KLT) [7], [8]. Among the existing transform domain techniques for speech enhancement in the literature, those based on DFT processing are usually favored in practical applications. This is due to several factors, such as lower computational complexity through the use of the fast Fourier transform (FFT), ease of implementation, ability to provide a trade-off between noise reduction and speech distortion at different frequencies, natural resemblance to the auditory processes taking place within human ear, and existence of efficient windowing techniques for the time-domain synthesis of the modified speech [9]. For all these reasons, the DFT-based methods, also known as frequency domain methods, have received much interest in the research community for more than three decades [10]–[12].

In these methods, the noisy speech spectrum is modified and then transformed back to the time domain to obtain the enhanced speech signal. However, in many applications such as mobile communication systems, the maximum algorithmic delay and the computational complexity are strictly limited. Moreover, use of the DFT is appropriate only for stationary signals, i.e., those with constant statistics over time. Yet, speech is known to be a quasi-stationary signal, i.e, one with approximately constant statistics over only short periods of time. For these reasons, in the frequency domain processing of speech signals, it is required to consider time segments of about 10-40 ms during which the statistics of the speech signal do not change significantly. This is realized by segmentation of the speech signal into short-time segments and subsequent processing of the Fourier coefficients of each segment individually. The processed coefficients across different time frames are inverse Fourier transformed and reassembled via overlap-add or overlap-save methods to produce the entire enhanced speech. This technique, referred to as short-time Fourier transform (STFT) processing, now serves as the basis to implement mainly all frequency domain methods of speech enhancement [9]. Besides being computationally efficient, this processing structure can handle different frequencies independently, which gives an appealing flexibility in exploiting the noise statistics and using our knowledge of speech perception to optimize the enhancement performance. As a result, most efforts in speech enhancement in the past have been devoted to this framework [7].

Assuming that the noise process is additive and that the noise and speech processes are (statistically) independent, many conventional methods in the STFT domain seek to estimate the speech DFT coefficients in an optimal sense. Due to the complex nature of speech DFT coefficients, however, they can be represented in either the real-imaginary or the amplitude-phase (polar) forms. In this regard, two broad types of methods can be recognized in the STFT domain: those attempting to separately estimate the real and imaginary components, and those aiming at the estimation of amplitude and/or phase of the clean speech DFT coefficients. Whereas the former are based on the assumption that the real and imaginary components of the DFT coefficients are independent, the latter assumes the amplitude and phase are independent components. Still, under a complex Gaussian model for speech DFT coefficients, it can be proved that these two assumptions are equivalent [13].

Considering the polar representation of complex DFT coefficients of speech signals, both the phase and the amplitude components are generally unknown. However, since the joint estimation of speech amplitude and phase can be mathematically challenging within a statistical optimization framework, a possible solution is to estimate each component separately and then combine them to produce the complex speech coefficients. In this regard, the spectral amplitude has been found to be perceptually more relevant than the spectral phase in the speech enhancement literature. According to the various experiments in [14], [15], the use of accurate estimates of speech phase, as compared to the noisy phase (i.e., that of the noisy speech), does not considerably improve the noise reduction performance. Furthermore, it was proved in [16] that the optimal estimate of speech DFT phase in the minimum mean square error (MMSE) sense is in fact the degraded noisy phase. For this reason, the majority of the efforts on the development of STFT-based noise reduction algorithms have focused on

the estimation of the speech spectral amplitude, also known in this context as short-time spectral amplitude (STSA). The most well-known methods for the estimation of speech spectral amplitude can be categorized into spectral subtraction [17], Wiener filtering [18] and statistical model-based methods [19]. The latter group of methods, also known as Bayesian STSA estimators, has actually been developed to overcome the disadvantage of the first two groups, which do not provide an optimal estimation of the STSA of the clean speech. In essence, Bayesian estimation of the speech STSA was first introduced by Ephraim and Malah in their seminal paper [16]. Therein, an MMSE-optimal estimator of the STSA is formulated and subsequently shown to achieve superior performance in enhancement when compared to other existing methods at this time. Following this groundbreaking work, several improved STSA estimators were suggested later in this direction, e.g. [20]–[22].

Under general conditions, a finite duration (one-dimensional) signal can be reconstructed (up to a scale factor) using only the phase of its DFT coefficients [23]. Therefore, in the context of speech enhancement, it may seem possible to first estimate the spectral phase more accurately and then attempt to reconstruct the signal from the phase information. But unfortunately, the accuracy of the reconstructed speech signal is extremely sensitive to the accuracy of the phase estimate, and such a technique for speech enhancement would require the ability to estimate the spectral phase very accurately, which is not an easy task [24]. Despite this fact, in recent years, there has been growing interest in the investigation of the effect of phase estimation in speech quality enhancement [25] and a few methods for the restoration of the spectral phase and its combination with the STSA have been suggested, e.g. [26]–[28].

The remainder of this paper is organized as follows. A brief background on various STFT-based noise reduction methods, considering both single- and multi-channel approaches, is presented first in Section II. In the case of single-channel, the major methods include spectral subtraction, Wiener filtering, MMSE and maximum *a posteriori* (MAP) estimators of the speech STSA, as well as the estimators of complex DFT (i.e., STFT coefficients). In the case of multi-channel, a selection of key beamforming approaches as well as post-filtering methods, (i.e. single-channel methods suitable to be applied on the output of a beamformer) are briefly discussed. Next, we briefly review the conventional noise estimation methods, which are crucial to the performance of noise reduction. At the end of this section, a theoretical review on performance assessment methods and common evaluation measures in speech enhancement is given. More recent advances on noise reduction methods in the STFT domain are presented in Section III. This includes the development of new speech priors used in the STSA estimators, recent advances in the estimation of noise power spectral density (PSD), speech enhancement in the short-time modulation domain, and the estimation of speech spectral phase. Furthermore, as one of the most important and challenging problems in the multi-channel case, recent methods for the estimation of the noise PSD matrix are also reviewed. Section IV includes a concise summary of this paper followed by important conclusions.

II. Background

In this section, we review in brief the conventional literature on single- and multi-channel methods developed for noise reduction in the STFT domain. This helps to state out the general problem, and to understand the motivations behind further developments in this area, as discussed in further details in Section III. Due to their difference in applications and processing strategies, we categorize and present the STFT domain methods based on the number of microphones (channels) used for the acquisition of noisy speech, i.e., single- and multi-channel.

A. Single-Channel Approaches

Despite their inherent performance limits and imposing distortion on the original speech signal, single-channel approaches are still an ongoing area of research in speech enhancement. For this reason, we present an overview of the conventional single-channel noise reduction methods in this section, which also provides a baseline for Section III. Assuming that the noise-corrupted speech, $x(t)$, consists of the clean speech, $s(t)$, and the additive noise, $\nu(t)$, we can write

$$x(t) = s(t) + \nu(t) \tag{1}$$

After sampling $x(t)$, the STFT of the resulting discrete-time signal, $x(\ell)$, can be implemented by segmentation of $x(\ell)$ into overlapping frames, multiplying the frames by a proper analysis window function and then taking DFT of each frame, as the following [7]

$$X(k,l) = \sum_{\ell=0}^{K-1} x(\ell + lZ)\, w(\ell)\, e^{-j2\pi\ell k/K} \tag{2}$$

where $w(\ell)$ is the window function, K is the frame length in samples, Z is the frame shift in samples, and k and l respectively denote the frequency bin and time frame indices. Typically, a Hamming window function can be used for $w(\ell)$ and a frame length of 20-40 ms along with a frame overlap (i.e., the ratio Z/K) of 50% or 75% are employed to implement the STFT analysis [7]. Invoking the additive noise model in (1), we obtain the following equivalent expression in the STFT domain

$$X(k,l) = S(k,l) + V(k,l) \tag{3}$$

where $S(k,\ell)$ and $V(k,\ell)$ respectively denote the STFT coefficients of $s(t)$ and $v(t)$. Assuming independence between the clean speech and noise, as well as independence across different frequency bins and time frames, the goal of noise reduction is to provide an estimate of the STFT of clean speech, denoted as $\hat{S}(k,l)$, which is as close as possible to the clean speech.

1) Spectral Subtractive Methods

Spectral subtraction is one of the first category of algorithms proposed for noise reduction in the frequency domain [7]. It is based on the simple principle that, having an estimate of the noise spectrum, $\hat{V}(k,l)$, an estimate of the clean speech spectrum, $\hat{S}(k,l)$, can be attained by subtracting the noise estimate from the noisy speech spectrum, $X(k,l)$. More specifically, assuming the similarity between the phase of the noisy speech and that of the clean speech, it follows that [17]

$$\hat{S}(k,l) = [\,|X(k,l)| - |\hat{V}(k,l)|\,]e^{j\Theta_X(k,l)} \tag{4}$$

where $|\cdot|$ denotes the amplitude and $\Theta_X(k,l)$ is the phase of $X(k,l)$. Note that the effect of noise on the clean speech phase is assumed negligible in (4), whereas in practice, availability of the clean speech phase or a better estimate of it to replace $\Theta_X(k,l)$ can provide further quality improvements [29]. Due to the inaccuracy in the noise estimate, $\hat{V}(k,l)$, the subtractive term, $|X(k,l)| - |\hat{V}(k,l)|$, can take on negative values and a half-wave rectification is conventionally used to mitigate this effect. This rectification causes a phenomenon known as musical noise, which can significantly degrade the speech quality up to a high degree [7]. This issue has been one of the main motives to develop more advanced spectral subtractive methods in the past, e.g., [30]–[32].

In practice, since the majority of noise estimation methods seek to estimate the noise spectral variance, $\sigma_v^2(k,l)$, defined as $E\{|V(k,l)|^2\}$, spectral subtractive methods are often formulated in the power domain rather than in the amplitude domain. In this regard, an estimate of the clean speech amplitude, $|\hat{S}(k,l)|$, can be obtained as

$$|\hat{S}(k,l)|^2 = |X(k,l)|^2 - \hat{\sigma}_v^2(k,l) \tag{5}$$

where $\hat{\sigma}_v^2(k,l)$ is an estimate of the noise spectral variance or the so-called PSD [7]. It is evident that the performance of spectral subtractive methods is highly controlled by the precision in the estimation of the noise PSD, $\sigma_v^2(k,l)$. Since the estimated speech amplitude can be written as a linear function of the noisy speech amplitude, it is often preferred to express spectrum estimation techniques in terms of a gain function. In this sense, the gain function for the estimator in (5) can be written as

$$G(k,l) \triangleq \frac{|\hat{S}(k,l)|}{|X(k,l)|} = \sqrt{1 - \frac{\hat{\sigma}_v^2(k,l)}{|X(k,l)|^2}} \tag{6}$$

For a better understanding of the concept of spectral subtraction, a block diagram of this method in its basic form is shown in Fig. 2. It is observed that, within this framework, only the spectrum amplitude is enhanced and the spectral phase is left unchanged.

One of the most important advances in the area of spectral subtractive methods is the use of masking properties of the human auditory system firstly introduced in [33]. The masking properties are essentially modelled by a noise masking threshold below which a human listener tolerates additive noise in the presence

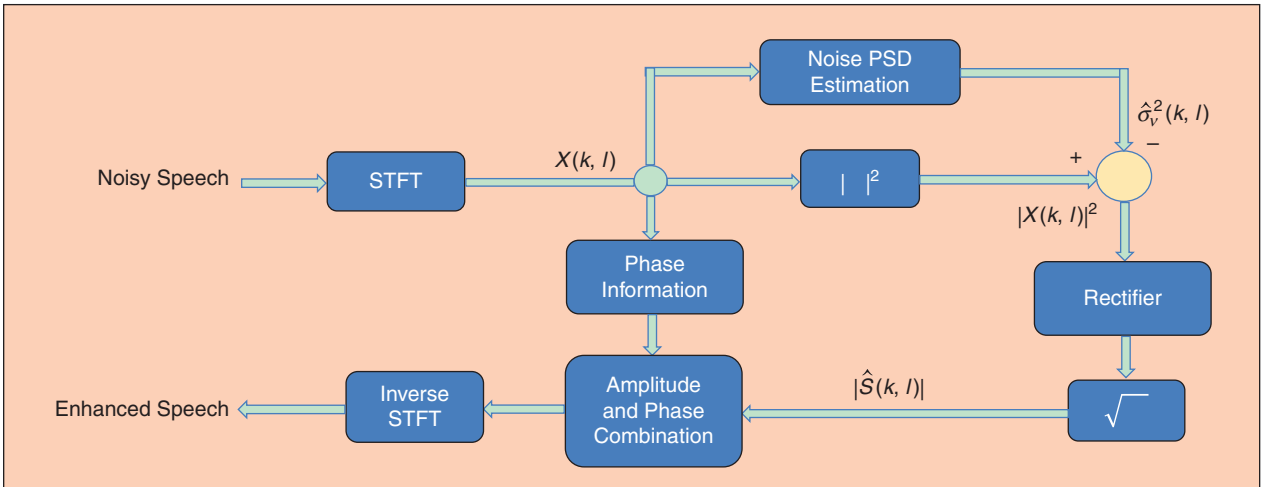

Figure 2. Block diagram of the basic spectral subtraction algorithm.

of speech [34]. In the generalized spectral subtractive methods, e.g. [35]–[36], there exist parameters which control the trade-off between the amount of noise reduction, the speech distortion and the residual musical noise. In [33], a few schemes are proposed based on the noise masking threshold in order to adjust the subtractive parameters in a perceptual sense. Therein, through the study of speech spectrograms as well as subjective listening tests, it is proved that the resulting enhanced speech is more pleasant to a human listener than without adaptive adjustment of the subtractive parameters.

The spectral subtraction algorithms are computationally simple to implement and fast enough for real-time applications. Nevertheless, the subtractive rules are based on the incorrect assumption that the cross terms between the clean speech and the noise are zero. In other words, considering (5) and the fact that $\hat{\sigma}_v^2(k,l)$ is used for $|V(k,l)|^2$, the speech squared amplitude $|S(k,l)|^2$ is not accurately equal to $|X(k,l)|^2-|V(k,l)|^2$, and the cross terms between the speech and noise have to be considered in the subtraction rule. In [37], a geometric approach (as opposed to the statistical approaches) to spectral subtraction is proposed that addresses this shortcoming of the spectral subtraction method. In that work, the phase difference between the clean speech and noise is exploited in order to obtain the spectral subtraction rule as a gain function. The resulting gain function depends on two key parameters, that is the *a priori* SNR and the noise PSD, and it possesses similar properties to those of the MMSE STSA estimator presented in [16] which will be further discussed in subsection II-A3 below. It is further shown through objective evaluations that the geometric algorithm performs significantly better than the traditional spectral subtraction algorithm under various conditions.

Other main contributions to the spectral subtraction method in the literature include spectral subtraction using oversubtraction [38], nonlinear spectral subtraction [39], multi-band spectral subtraction [40], MMSE-based spectral subtraction [35], extended spectral subtraction [41], use of adaptive gain averaging [31] and selective spectral subtraction [42]. Even though spectral subtraction is one of the oldest methods of noise reduction in the STFT domain, there still exists ongoing research on this topic.

2) Wiener Filtering Based Methods

The spectral subtractive methods discussed in the previous section are based on the heuristic assumption that one can obtain an estimate of clean speech spectrum by subtracting the estimated noise spectrum from the observations spectrum. Despite being intuitively pleasing and computationally simple, this method cannot make any claim of optimality. In this part, we briefly review the concept of Wiener filtering in the STFT domain. In this approach, the estimated speech spectrum is obtained as $\hat{S}(k,l) = W(k,l)X(k,l)$ where $W(k,l)$ denotes the corresponding gain function. The latter is derived by minimizing the mean square error (MSE) between the clean and estimated speech spectra, which is mathematically expressed as

$$\hat{W}(k,l) = \underset{W}{\operatorname{argmin}}\ E\{|S(k,l) - WX(k,l)|^2\} \tag{7}$$

with $E\{.\}$ denoting the statistical expectation. Solving the above, we obtain the general form of the complex-valued Wiener filter gain as

$$\hat{W}(k,l) = \frac{\sigma_{sx}(k,l)}{\sigma_x^2(k,l)} \tag{8}$$

where $\sigma_{sx}(k,l)$ denotes the cross-PSD between the clean and noisy speech defined as $E\{S(k,l)X^*(k,l)\}$ and $\sigma_x^2(k,l)$ denotes the noisy speech PSD [43]. In practice, both σ_{sx} and σ_x^2 in (8) are unknown and have to be estimated. Henceforth, we may drop the time frame and frequency indices for improved readability. Even though the estimation of σ_x^2 can be done in a straightforward way, such as recursive smoothing of the observations, $X(k,l)$, estimation of the cross-term σ_{sx} is generally challenging and depends on the application [44]. Assuming uncorrelated clean speech and noise signals, σ_{sx} and σ_x^2 respectively simplify to the clean speech PSD, σ_s^2, and the sum $\sigma_s^2+\sigma_v^2$. Now, by defining the *a priori* SNR as $\zeta = \sigma_s^2/\sigma_v^2$, the Wiener filtering gain can be expressed as $\zeta/(1+\zeta)$. The *a priori* SNR, which is a critical parameter in the context of noise reduction, can be estimated through the conventional decision-directed approach [16] and its more advanced variations found in [45]–[47]. The aforementioned method is the most conventional way for computing the Wiener filter gain function from the available noisy speech. Several alternative methods have been proposed in the relevant literature in order to implement the Wiener filter, which are summarized in Table 1.

The latest category in Table 1, i.e. the codebook-based method, is known to perform better in the presence of highly non-stationary noise, while eliminating the need to employ noise estimation algorithms. Yet, its main shortcoming is the complexity arising from the required search for an optimal vector of linear prediction coefficients in a possibly large dimensional codebook, as well as the need to generate the codebook *a priori* with the help of training data. Another interesting research avenue in the field of Wiener filtering is the application of psychoacoustics in order to introduce additional constraints in the design of the Wiener filtering gain. As such, the masking properties of human auditory system have been employed to determine the thresholds on the speech or noise distortion introduced by Wiener filtering [49], [50].

Table 1.
Main Wiener filtering methods in the STFT domain.

Method	Filtering Gain	Properties
Square-root Wiener filter [7]	$\sqrt{\frac{\sigma_{sx}(k,l)}{\sigma_x^2(k,l)}}$	PSD of the enhanced speech is theoretically identical to that of the clean speech.
Parametric Wiener filter [43]	$\left(\frac{\sigma_s^2(k,l)}{\sigma_s^2(k,l)+\alpha\sigma_v^2(k,l)}\right)^{\beta}$	The use of parameter β allows to compromise between noise reduction and speech distortion
Iterative Wiener filtering [43]	At iteration i, the estimated speech is used to estimate the speech PSD and therefore the Wiener filtering gain, as following: $\hat{S}_{i+1} = W_i X \ \hat{S}_{i+1} \rightarrow W_{i+1}$	More precise estimate of the Wiener filtering gain compared to the non-iterative version, if convergence occurs
Constrained Wiener filtering [7]	$\frac{1}{1+\sqrt{\frac{\sigma_v^2}{\sigma_s^2}}}$	Allows a compromise between the amount of speech distortion and noise distortion
Codebook-driven Wiener filtering [48]	$\frac{\frac{\hat{\sigma}_s^2}{\lvert\hat{D}_s(k,l)\rvert^2}}{\frac{\hat{\sigma}_s^2}{\lvert\hat{D}_s(k,l)\rvert^2}+\frac{\hat{\sigma}_v^2}{\lvert\hat{D}_v(k,l)\rvert^2}}$	Use of auto-regressive spectrum models estimated from speech/noise codebooks

3) MMSE-Based (Bayesian) Estimators of STSA

In this subsection, we discuss an important category of STSA estimators, which are optimal in the amplitude MMSE sense. Basically, the Bayesian STSA estimation problem can be formulated as the minimization of the expectation of a cost function representing a measure of distance between the true and the estimated clean speech STSAs, denoted respectively by $A(k,l)$ and $\hat{A}(k,l)$. This problem can be expressed as

$$\hat{A}^{(o)} = \underset{\hat{A}}{\operatorname{argmin}}\ E\{C(A,\hat{A}) \mid X\} \tag{9}$$

where $C(.)$ is a particular Bayesian cost function and $\hat{A}^{(o)}$ is the optimal STSA estimate. Similar to the spectral subtractive methods discussed earlier, the STSA estimate is combined with the noisy phase of speech to provide an estimate of speech STFT coefficients. This approach was firstly established by Ephraim and Malah in [16] wherein the cost function $C(.)$ was taken as the squared error between $A(k,l)$ and $\hat{A}(k,l)$, implying $\hat{A}^{(o)}$ to be the MMSE estimate of the speech STSA. Further proceeding with (9) requires the knowledge of the distribution of speech STSA conditioned on observation, i.e. $p(A|X)$, since

$$\begin{aligned} E\{C(A,\hat{A})\} &= \int\int C(A,\hat{A})\ p(A,X)dA\,dX \\ &= \int\left[\int C(A,\hat{A})p(A|X)dA\right]p(X)dX \end{aligned} \tag{10}$$

where actually the term inside the brackets has to be minimized with respect to $\hat{A}$. This has been conventionally done in a Bayesian framework for $p(A|Y)$ under the assumption that the noise coefficients, V, follow a zero-mean complex circularly symmetric Gaussian distribution, and as a result, the speech spectral phase, Ω, follows a uniform distribution and the speech STSA has a Rayleigh distribution. Moreover, speech phase and amplitude are supposed to be independent. Under these assumptions, it follows that [16]

$$\begin{aligned} p(X \mid A,\Omega) &= \frac{1}{\pi\sigma_v^2}\exp\left(-\frac{\lvert X - Ae^{j\Omega}\rvert^2}{\sigma_v^2}\right) \\ p(A,\Omega) &= p(A)\ p(\Omega) = \frac{A}{\pi\sigma_s^2}\exp\left(-\frac{A^2}{\sigma_s^2}\right) \end{aligned} \tag{11}$$

Based on these assumptions and considering the cost function $C(A,\hat{A})$ to be $(A-\hat{A})^2$, the MMSE estimator of the speech STSA has been derived as a closed-form solution in [16]. Although this STSA estimator provided considerable improvements with respect to the previous spectral subtractive or Wiener-based methods, it did not take into account the most subjectively meaningful Bayesian cost function. Based on this fact, the same authors suggested a logarithmic version of their MMSE estimator in [20] where the cost function exploits the log-spectra of the clean and estimated STSA. Therein, it was shown that the log-spectra is more suitable as the distortion measure and further improvements with respect to the original MMSE estimator were achieved in most experiments. Later, Loizou in [21] introduced the idea of perceptually (to human ear) motivated cost functions and derived STSA estimators that emphasize on the spectral peak (formants) information and STSA estimators which take into account the auditory masking effects of the human audition system.

Therein, he proposed three classes of Bayesian estimators. The first class of the estimators emphasizes spectral peak information of the speech signal, the second class uses a weighted Euclidean cost function that takes into account the aforementioned auditory masking effects and the third class of estimators is developed to account for spectral attenuation. It was concluded that, out of the three classes of the suggested Bayesian estimators, those based on the auditory masking effect perform best in terms of having less residual noise in the enhanced speech and better speech quality.

Within the same direction, another major class of Bayesian STSA estimators was proposed in [22], which is known as the β-order MMSE estimator. The corresponding cost function involves a parameter named β and employs β powers of the amplitude spectra. Thanks to the degree of freedom provided by this parameter, trade-offs between the amount of noise reduction and speech distortion were achieved therein and a few schemes for the experimental or adaptive selection of this parameter were contributed. The experimental results proved the advantage of the namely β-SA estimator, as compared to the previous versions of STSA estimation. Along the same direction, later in [51], it was proposed to exploit a spectrally weighted development of the β-order MMSE cost function including a new weighting parameter called a. Therein, new psychoacoustical schemes were suggested for the selection of the two parameters, i.e. $a$and β, based on the properties of human auditory system. Performance evaluations revealed improvements in the so-called Wβ-SA estimator with respect to using the previously suggested MMSE cost functions in this field. Later in [52], a more generalized Bayesian cost function was introduced by involving a new spectral weighting term and it was indicated that the resulting STSA estimator, named as generalized weighted SA (GWSA), provides further flexibility in the adjustment of the STSA gain function. All the aforementioned STSA estimators can actually be derived as a particular case of the latter.

To facilitate the discussion of the conventional Bayesian STSA estimators with the underlying cost functions, a summary of the major STSA estimators is indicated in Table 2. In this table, γ is the *a posteriori* SNR defined as $|X|^2/\sigma_v^2$, the gain function parameter υ is $\zeta\gamma/(1+\zeta)$ and $M(.,.;.)$ denotes the confluent hypergeometric function. Note that p, β and a are parameters that shape the STSA gain function, and as explained, a few efficient schemes for their determination have been proposed in the references in Table 2.

Table 2.
Major Bayesian estimators of speech STSA.

Method	Bayesian Cost Function	Gain Function	Properties
MMSE [16]	$(A-\hat{A})^2$	$\frac{\sqrt{\nu}}{\gamma}\Gamma(1.5)\,M(-0.5,1;-\nu)$	Basic version of Bayesian STSA estimators, optimal in the amplitude MMSE sense
Log-MMSE [20]	$(\log A_k - \log \hat{A}_k)^2$	$\frac{\nu}{\gamma}\exp\left(\frac{1}{2}\int_{\upsilon}^{\infty}\frac{e^{-t}}{t}dt\right)$	Outperforms the basic version through the use of the logarithmic distortion measure (cost function) for speech
WCOSH [21]	$A^p\left(\frac{A}{\hat{A}}+\frac{\hat{A}}{A}-1\right)$	$\frac{\sqrt{\nu}}{\gamma}\sqrt{\frac{\Gamma\left(\frac{p+3}{2}\right)M\left(-\frac{p+1}{2},1;-\nu\right)}{\Gamma\left(\frac{p+1}{2}\right)M\left(-\frac{p-1}{2},1;-\nu\right)}}$	Weighted cosine hyperbolic cost function, a symmetric distortion measure exploiting auditory masking effects
WE [21]	$A^p(A-\hat{A})^2$	$\frac{\sqrt{\nu}}{\gamma}\frac{\Gamma\left(\frac{p+1}{2}+1\right)}{\Gamma\left(\frac{p}{2}+1\right)}\frac{M\left(-\frac{p+1}{2},1;-\nu\right)}{M\left(-\frac{p}{2},1;-\nu\right)}$, $p>-2$	Distortion measure motivated by the perceptual weighting technique used in low-rate analysis-by-synthesis speech coders
β-SA [22]	$(A^\beta-\hat{A}^\beta)^2$	$\frac{\sqrt{\nu}}{\gamma}\left[\Gamma\left(\frac{\beta}{2}+1\right)M\left(-\frac{\beta}{2},1;-\nu\right)\right]^{1/\beta}$, $\beta>-2$	Motivated first by the generalized spectral subtraction method, provides gain function adjustments by the selection of parameter β
$W\,\beta$-SA [51]	$\left(\frac{A^\beta-\hat{A}^\beta}{A^\alpha}\right)^2$	$\frac{\sqrt{\nu}}{\gamma}\left(\frac{\Gamma\left(\frac{\beta-2\alpha}{2}+1\right)M\left(-\frac{\beta-2\alpha}{2},1;-\nu\right)}{\Gamma(-\alpha+1)\,M(\alpha,1;-\nu)}\right)^{1/\beta}$ $\beta>2(\alpha-1)$, $\alpha<1$	Further flexibility in the gain function, selection of parameters α and β based on psycho-acoustical properties of human audition

4) MAP Estimators

In [10], the first statistical-based estimator of the speech STSA was proposed in the form of an ML estimator and a few closed-form solutions were derived, based on the following

$$\hat{A}^{(ML)} = \underset{A}{\operatorname{argmax}}\ E_{\Omega}\{p(X \mid A, \Omega)\} \tag{12}$$

where E_{Ω}. denotes the expectation over speech phase, Ω. Yet, apart from the limited performance, an ML estimator does not take into account the distribution of speech STSA prior, whereas a proper model for the speech prior can be considered in the MAP estimation. In [53], under a complex Gaussian assumption for the speech prior and a Bayesian framework, MAP estimators of the speech STSA were derived as simpler alternatives to the Ephraim and Malah's MMSE-based approach. Therein, three different estimators were proposed, namely, the joint MAP estimator of speech spectral amplitude and phase, the MAP estimator of the speech spectral amplitude and the MMSE estimator of speech PSD. The joint MAP spectral amplitude and phase estimator can be expressed as [53]

$$(\hat{A}^{(\mathrm{MAP})}, \hat{\Omega}^{(\mathrm{MAP})}) = \underset{A,\Omega}{\operatorname{argmax}}\ p(A, \Omega \mid X) \tag{13}$$

and closed-form solutions for the speech spectral amplitude and phase are derived from (13). The interesting result, however, is that the estimator of the speech spectral phase obtained by (13) is just the noisy phase of speech observations. The same result was deduced in [16] with the MMSE estimate of speech spectral phase. Next, the spectral amplitude-only estimator can be given by solving the following [53]

$$\hat{A}^{(\mathrm{MAP})} = \underset{A}{\operatorname{argmax}}\ E_{\Omega}\{p(A, \Omega \mid X)\} \tag{14}$$

where, using an exponential approximation to the Rician distribution obtained from $E_{\Omega}\{p(A, \Omega \mid X)\}$, leads to a closed-form solution. It is shown that this solution is a generalized form of the approximate solution to the ML estimator proposed in [10]. Next, by deriving an expression for the second moment of the Rician posterior, i.e. $E\{A^2 \mid X\}$, which is actually the MMSE estimate of the speech spectral variance, σ_s^2, and taking its square root, an estimate of speech spectral amplitude is obtained and combined with the noisy phase. Analysis of the behavior of the corresponding gain functions for all three estimators shows that they have a similar performance to the Ephraim and Malah's solution, whilst they permit a more straightforward implementation and simpler expressions by avoiding Bessel and Hypergeometric functions.

More recently in [54], it was indicated through extensive experimentations that the class of super-Gaussian distributions fits speech STSA priors more properly than the conventional Rayleigh deduced from the complex Gaussian assumption for speech STFT coefficients. Therein, within the framework of MAP spectral amplitude estimators, the distribution of the speech spectral amplitude is modeled by a simple parametric function, which allows a high approximation accuracy for Laplace- or Gamma-distributed real and imaginary parts of the speech STFT coefficients. Also, the statistical model can be adapted using the noisy observations to optimally fit the distribution of the speech spectral amplitudes. Based on the super-Gaussian statistical model, two computationally efficient MAP spectral amplitude estimators are derived, which outperform the previously proposed ones in [53] while owning the same simplicity as the estimators in [53]. The two estimators in [54] include a joint amplitude-phase estimator and an amplitude-only estimator and can be both expressed as extensions of the MAP estimators proposed in [53]. In Table 3, a summary of the different MAP estimators for the spectral amplitude is presented.

5) Estimators of Complex STFT

Considering the complex-valued STFT coefficients of speech, one can tend to estimate in the rectangular form the real and imaginary parts. In [13], [55], such estimators of complex-valued speech STFT coefficients have been proposed using different distributions for speech prior and noise. Therein, in order to derive closed-form solutions, it is assumed that the real and imaginary parts of the complex speech STFT coefficients are independent, which can be thought of as a counterpart to the independence of the spectral amplitude and phase assumed in the speech amplitude estimators. In fact, under a complex Gaussian distribution for the speech STFT prior, both assumptions are equivalent, though in general, they are not the same [13].

According to the evaluations presented in [13], the STFT coefficients of short-time stationary clean speech signals can be better modelled by a super-Gaussian density such as the two-sided exponential (Laplace) and the two-sided Gamma density. Therefore, the MMSE estimation of the complex-valued speech STFT coefficients has been handled in [13], [55] under certain super-Gaussian distributions. Assuming independent real and imaginary parts for speech coefficients, this problem leads to the following

$$\hat{S} = E\{S \mid X\} = E\{S_R \mid X_R\} + jE\{S_I \mid X_I\} \tag{15}$$

where the subscripts R and I denote the real and imaginary parts, respectively. It can be observed from (15) that the estimation of complex-valued speech coefficients is

Table 3.
MAP estimators of speech spectral amplitude.

Method	Gain function
Joint MAP estimator of speech spectral amplitude and phase [53]	$\frac{\zeta+\sqrt{\zeta^2+2(1+\zeta)(\zeta/\gamma)}}{2(1+\zeta)}$
MAP estimator of speech spectral amplitude [53]	$\frac{\zeta+\sqrt{\zeta^2+(1+\zeta)(\zeta/\gamma)}}{2(1+\zeta)}$
MMSE estimation of speech spectral variance (PSD) [53]	$\sqrt{\frac{\zeta}{1+\zeta}\frac{1+\nu}{\gamma}}$
Joint MAP estimator of speech spectral amplitude and phase (super-Gaussian speech spectral amplitude) [54], with a and b as the super-Gaussian parameters	$U+\sqrt{U^2+\frac{a}{2\gamma}}, U=\frac{1}{2}-\frac{b}{4\sqrt{\gamma\zeta}}$
MAP estimator of speech spectral amplitude (super-Gaussian speech spectral amplitude) [54]	$U+\sqrt{U^2+\frac{a-1/2}{2\gamma}}, U=\frac{1}{2}-\frac{b}{4\sqrt{\gamma\zeta}}$

in fact decomposed into the independent estimation of their real and imaginary parts. In this sense, considering the same distribution for S_R and S_I, we arrive at the same estimators for the real and imaginary components of speech STFT coefficients [13], [55]. Further proceeding with (15) requires assuming distributions for noise and speech STFT coefficients. The complex Gaussian distribution has been conventionally used to model the noise coefficients in the literature. However, different super-Gaussian distributions have been exploited to model the real and imaginary parts of the speech STFT coefficients in [13], as the following:

$$\text{Gaussian: } p(S_R)=\frac{1}{\sqrt{\pi\sigma_s}}\exp\left(-\frac{S_R^2}{\sigma_s^2}\right)$$

$$\text{Laplacian: } p(S_R)=\frac{1}{\sigma_s}\exp\left(-2\frac{|S_R|}{\sigma_s}\right)$$

Two-sided Gamma (special case):

$$p(S_R)=\frac{\sqrt[4]{3/2}}{2\sqrt{\pi\sigma_s|S_R|}}\exp\left(-\frac{\sqrt{3/2}|S_R|}{\sigma_s}\right) \quad (16)$$

where σ_s^2 is the speech spectral variance defined by $E\{|S|^2\}$. The same distributions are used for the imaginary part, S_I. In [55], two special cases of the two-sided generalized Gamma distribution were also considered for the real and imaginary parts of the speech prior. The mathematical expressions for the corresponding estimators, however, are more tedious than those for the Bayesian estimators in Table 2 and requires more computational burden. Similar estimators were also derived in [13] under the complex Laplacian distribution for noise coefficients, yet, in addition to the further complexity involved, no performance improvements were reported with respect to the estimators with the complex Gaussian noise distribution.

Considering the noise reduction performance of the complex STFT estimators, it is concluded in [13], [55] that they perform slightly better than the MAP estimators of the speech spectral amplitude. Also, employing super-Gaussian distributions for the speech prior led to improvements compared to the conventional Gaussian. However, comparing the complex STFT estimators to the Bayesian estimators of the speech spectral amplitude (namely STSA estimators), it was revealed that the latter performs as well or better than the former. According to the experimentations in [13] and references therein, the reason is that speech amplitude and phase are indeed statistically less dependent than the real and the imaginary parts of speech complex coefficients. Note that the independence between amplitude and phase or between real and imaginary components of speech STFT has to be assumed in order to have a mathematically tractable solution for all speech estimators. Also, from a computational standpoint, since the speech STSA estimation requires the computation of only one estimate, it is overly computationally less complex than the estimators of the complex STFT.

B. Multi-Channel Approaches

In this section, we present a brief overview of the most famous multi-channel noise reduction approaches, conventionally known as beamforming techniques. In this regard, a few beamforming methods and their variations have been proposed and widely used in the speech enhancement literature. It should be noted that, compared to other applications such as radar and sonar signals, implementation of the beamforming algorithms for speech signals has shown to be more challenging. This

is because speech is typically a wideband signal and it owns the features of highly non stationary signals. Also in practical environments, background noise or interference may own the same spectral features as those of the clean speech signal [3].

Suppose that we have an array consisting of N microphones capturing a desired source of speech. The microphone observations are contaminated by additive noise which, in general, can be correlated across different microphones. Therefore, the received signals can be expressed in the vector form as the following

$$\mathbf{X}(k,l) = S(k,l)\mathbf{A}(k) + \mathbf{V}(k,l) \tag{17}$$

where $\mathbf{X}(k,l) = [X_1(k,l),\ X_2(k,l),...,\ X_N(k,l)]^T$ and $\mathbf{V}(k,l) = [V_1(k,l), V_2(k,l),...,V_N(k,l)]^T$ respectively denote the set of observations and noises received by the microphone array; and $\mathbf{A}(k)$ is the so-called steering vector which depends on the direction of arrival (DOA) of the speech source with respect to the microphone array. Considering the estimation of the parameter DOA and so the steering vector, $\mathbf{A}(k)$, a few major techniques such as ML methods, subspace-based methods, using beamforming and compressive sensing approaches exist in the literature [56]–[57]. Here, we focus on a few major beamforming and post-filtering techniques that tend to estimate the clean speech spectrum under the independence of noise and speech, given that the DOA is known or estimated beforehand. In general, beamformers are actually linear filters in the STFT domain and can be represented by a weight vector, $\mathbf{W}(k,l)$, applied over the microphone array observations, $\mathbf{X}(k,l)$. Next, as shown in Fig. 3, the resulting single-channel output is fed into a linear post-filter that can be expressed through a gain function, $G_{\text{post}}(k,l)$; and the ultimate estimate of the speech STFT is given at the output of the post-filter.

1) Conventional Beamforming Techniques

Although single-channel noise reduction algorithms are generally able to improve the speech quality, with the advance in today's technology, use of microphone arrays has become more popular. In this regard, beamforming techniques, due to taking advantage of the spatial information across different microphones, enable further noise reduction without imposing considerable distortion on speech. In this section, the most well-known beamforming techniques in the frequency domain are briefly discussed. These techniques include delay-and-sum (DAS) beamformer, Wiener filter (multi-channel), distortionless Wiener filter, maximum SNR filter and minimum variance distortionless response (MVDR) beamformer [58]. For ease of notation, we may drop the indices k and l henceforth.

The simplest beamforming technique is the DAS beamforming which compensates the relative delay across the speech components received by different microphones and then sums up the delay-compensated observations to form the enhanced speech as $\mathbf{A}^H\mathbf{X}$. The simplicity in implementation and the small amount of imposed distortion on speech are the main advantages of this technique. Yet, due to the limited performance improvement achieved by this technique, often in practice, other beamformers are favored [58].

Similar to its single-channel version, the multi-channel Wiener filter is derived based on minimizing the MSE between the clean and estimated speech spectra, resulting in the following

$$\begin{aligned}\mathbf{W}^{(1)} &= \underset{\mathbf{W}}{\operatorname{argmin}}\ E\{\,|\,S - \mathbf{W}^H\mathbf{X}\,|\,^2\} \\ &= \frac{\Sigma_{\mathbf{VV}}^{-1}\,\Sigma_{\mathbf{SS}}\,\mathbf{u}}{1 + \operatorname{tr}\{\Sigma_{\mathbf{VV}}^{-1}\,\Sigma_{\mathbf{SS}}\}}\end{aligned} \tag{18}$$

where $\Sigma_{\mathbf{VV}}$ is the $N\times N$ noise spatial PSD matrix defined as $E\{\mathbf{V}\mathbf{V}^H\}$, the speech signal PSD matrix $\Sigma_{\mathbf{SS}}$ is given as $\sigma_s^2\mathbf{A}\mathbf{A}^H$, tr{.} denotes the matrix trace operation and $\mathbf{u}$ is the $N\times 1$ unitary vector defined as $[1,0,0,\ldots,0]^T$. Even though this technique is optimal in the MSE sense and is capable of providing a high level of noise reduction, it imposes considerable amount of distortion on the speech component. Also, the estimation of the noise PSD matrix, $\Sigma_{\mathbf{VV}}$, as well as the speech PSD, σ_s^2, is a challenging task in general [59]. By adding the constraint $\mathbf{W}^H\mathbf{A} = 1$, to the multi-channel Wiener filter in (18), a distortionless modification of this method can be obtained as the following [59]

$$\mathbf{W}^{(2)} = \frac{\Sigma_{\mathbf{VV}}^{-1}\,\mathbf{A}}{\mathbf{A}^H\,\Sigma_{\mathbf{VV}}^{-1}\,\mathbf{A}} \tag{19}$$

Note that the constraint $\mathbf{W}^H\mathbf{A} = 1$ ensures theoretically that the speech signal arriving at the DOA is passed through the beamformer without being distorted.

Instead of minimizing the MSE between the clean and estimated speech, another criterion to achieve the maximum possible noise reduction is to maximize the SNR of the output speech. This is achieved via the

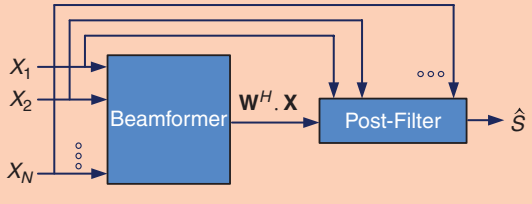

Figure 3. Typical microphone array speech enhancement system consisting of a beamformer and a post-filter.

maximum SNR spatial filter which can be expressed as the following [58]

$$\mathbf{W}^{(3)} = \frac{\mathbf{W}^{(1)}}{\sqrt{\mathbf{W}^{(1)^H}\mathbf{W}^{(1)}}} \tag{20}$$

It is observed that the maximum SNR filter in the above is in fact a normalized version of the multi-channel Wiener filter. Even though this beamformer is designed to achieve the most improvement in the SNR, there can be uncontrollable distortion in the enhanced speech signal.

The most famous beamformer in the literature is the MVDR technique which aims at minimizing the noise PSD in the output speech signal subject to a distortionless constraint on the speech. This leads to the following [58]

$$\mathbf{W}^{(4)} = \underset{\mathbf{W}}{\operatorname{argmin}}\ \mathbf{W}^H \Sigma_{\mathbf{VV}}^{-1}\mathbf{W},\ \ \mathbf{W}^H\mathbf{A} = 1 \tag{21}$$

which leads to the same expression as that for the distortionless Wiener filter in (19). In fact, minimizing the MSE and the noise PSD in the enhanced speech result in the same solution for the beamformer and thus, the MVDR beamformer is the distortionless version of the multi-channel Wiener filter. Even though in theory no distortion should be imposed on the speech component, in practice, due to the inaccuracy in the estimation of the steering vector, $\mathbf{A}$, and the noise PSD matrix, $\Sigma_{\mathbf{VV}}$, a minimum level of distortion is always inevitable. More elaboration and further insights into the MVDR beamformer can be found in [60].

2) Conventional Post-Filtering Techniques

In many cases, the performance gain yielded by a beamforming technique is not sufficient, however, it can be increased by properly adopting a post-filtering technique on the output of the beamformer. A post-filtering method is often employed to remove the non-coherent parts of the signal at beamformer's output, and as seen in Fig. 3, the transfer function of this post-filter is generally derived from the spatial cross-PSDs (or coherence functions) of the sensor signals [61]. These cross-PSDs should be either estimated empirically or derived from a specific model for the coherence function such as that for a diffuse noise field.

The first major post-filtering method in the STFT domain was proposed by Zelinski in [62] where a robust adaptive approach is introduced to obtained the post-filter gain based on the estimation of the cross-PSDs across microphone observations. The main benefits of his approach are that the noise PSD in microphone observations can be variable and that the speech output signal is theoretically free of musical tones or other residual noise remained after the beamforming. In [63], Zelinski's heuristic approach was formulated as a Wiener filter, i.e. in an MMSE sense, and then extended from $N = 4$ microphones to an arbitrary number of sensors.

In [61], the basic principles of conventional post-filtering along with their theoretical analysis are presented and conclusions has been drawn about behavior of Wiener-based post-filters in real environments. Also, post-filter transfer functions have been extended to an acoustic environment, i.e. that with room reverberation. Therein, one general form of the post-filter derived based on the Wiener-Hopf equation, is expressed as the following filter gain

$$W^{(\text{post})} = \frac{\frac{2}{N(N-1)}\Re\left\{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sigma_{\check{x}_i\check{x}_j}\right\}}{\sigma_y^2} \tag{22}$$

where $\Re\{.\}$ denotes the real value, σ_y^2 is the PSD of beamformer output, i.e. $y = \mathbf{W}^H\mathbf{X}$, and $\sigma_{\check{x}_i\check{x}_j}$ is the cross-PSD between the beamformer outputs $W_i^* X_i$ and $W_j^* X_j$. In [64], an approach for the estimation of the PSD terms in (22) is proposed based on the concept of noise field coherence. The latter is measured in terms of a complex coherence function that may take different forms, depending on the type of the noise field. As a common application, for a diffuse noise type, the coherence function is

$$\phi_{i,j} \triangleq \frac{\sigma_{\check{x}_i\check{x}_j}}{\sqrt{\sigma_{\check{x}_i}^2\sigma_{\check{x}_j}^2}} = \operatorname{sinc}\left(\frac{2\pi f d_{i,j}}{C}\right) \tag{23}$$

with $f = (k/K)(f_s/2)$ as the frequency, $d_{i,j}$ the spacing between sensors i and j, and C the sound velocity. Clearly, having estimates of the PSD terms $\sigma_{\check{x}_i}^2$ and $\sigma_{\check{x}_j}^2$, the spatial cross-PSD can be estimated by (23) using the proper coherence function.

Finally in [65], an efficient post-filtering algorithm for a special type of beamformer, namely, the generalized sidelobe canceller (GSC), is proposed. This type of beamformer is very useful in suppressing directional noise arriving from specific directions toward the microphone array, yet, its performance degrades to a large extent in the presence of diffuse non-stationary noise. The suggested post-filter takes advantage of the noise-only components constructed within the GSC structure and is able to deal with diffuse noise fields effectively. In [65] also, two single-channel approaches, the optimally-modified log-spectral amplitude (OM-LSA) in [66] and the mixture-maximum (MIXMAX) method in [67], have been adopted as post-filters by being concatenated to the beamformer output.

C. Estimation of Noise PSD

From the previous parts, it is evident that speech estimators in the STFT domain are generally a function of the noise PSD, $\sigma_v^2(k,l)$, either directly or indirectly through the *a priori* and *a posteriori* SNRs. Therefore, apart from the estimator type and speech prior model, the performance of the speech spectral estimators is heavily dependent on the accuracy of the given noise PSD estimate. Indeed, the noise PSD can be either underestimated of overestimated. On the one hand, the underestimation of the noise PSD leads to an under-suppression of noisy speech, and therefore, an unfavorably large amount of residual noise. On the other hand, the overestimation of the noise PSD generally results in an over-suppression of noisy speech, and consequently a potential loss in speech quality or intelligibility [68]. Noise PSD estimation is thus a crucial part of any noise reduction algorithm in the STFT domain, and is, in particular, challenging when speech is corrupted by non-stationary noise. In this part, we briefly overview three main approaches to noise PSD estimation, i.e., voice activity detection, minimum tracking and minimum controlled recursive averaging. Further, we discuss some of the more recent developments on each of these approaches in the literature.

1) Voice Activity Detection

The earliest noise PSD estimation methods exploit the fact that between durations of talker activity and even between word syllables, speech is absent for a short moment. During these moments, the noisy speech degenerates to the noise realization and thus the estimation of the noise PSD is feasible. To detect the non-speech segments, there exist numerous methods in the literature, which are often referred to as voice activity detection (VAD). Most VAD approaches rely on the fact that certain statistics, such as the energy or the log-energy of the noise-only process and the noisy speech process, are different. By comparing these statistics to the actual energy or log-energy of the signal, it can be decided whether speech is absent or present [69]. Let the two hypotheses $\mathcal{H}_0(k,l)$ and $\mathcal{H}_1(k,l)$ respectively indicate that speech is present and absent in a particular time frame and frequency bin. Then, a VAD-based approach estimates the noise PSD by recursively smoothing the noisy observations under $\mathcal{H}_0(k,l)$, as the following

$$\hat{\sigma}_v^2(k,l) = \begin{cases} \kappa\, \hat{\sigma}_v^2(k,l-1) + (1-\kappa)|X(k,l)|^2, & \mathcal{H}_0(k,l) \\ \hat{\sigma}_v^2(k,l-1), & \mathcal{H}_1(k,l) \end{cases} \tag{24}$$

where $0 \le \kappa \le 1$ is a smoothing constant. Although VADs are conceptually simple, their capability to accurately estimate the noise PSD and track fast changes in non-stationary noise is limited. For this reason, there has been various developments and their modifications of this approach proposed in the literature.

Apart from the conventional statistical model-based VADs such as [69], characteristics of speech and non-speech segments have also been modelled by hidden Markov models (HMMs). For example, in [70], a decision-tree algorithm that tends to combine the scores of HMM-based speech/non-speech models and speech pulse information is employed in order to reject far-field speech for speech recognition systems. In this work, the state duration is controlled by the state-transition probabilities of the HMMs and speech pulse information. Also, in [71] and [72], proper statistical models are used to characterize speech and non-speech signals, using decision logics governing the switching between speech and non-speech states in the HMMs. Yet, in the Gaussian mixture model (GMM) based VAD of [71], state duration is governed by the number of speech frames detected by the GMMs in a fixed-length buffer, while in the GMM-VAD of [72] state duration is governed by a hangover and hand-before scheme which detects the consonants occurred at the beginning, middle and the end of speech segments. Note that both HMM and GMM-based VADs require ground-truth speech/non-speech segments for training their statistical models. However, in the GMM-based VAD of [73], the need for a training stage has been eliminated by applying speech enhancement as a pre-processing step to improve the SNR of noisy speech segments. This enables the VAD algorithm to use either log-likelihood ratio tests or the comparison with energy-based thresholds in order to discriminate speech/non-speech segments.

The concept of using noise reduction to improve the detection performance of VADs in low SNR conditions has also been explored in [74]. Therein, the basic idea is to use features extracted from a noise-reduced representation of the original noisy speech via using non-negative sparse coding. In this regard, the speech STSA is decomposed on a speech dictionary learned from clean speech data and a noise dictionary learned from noise samples. Next, the coefficients corresponding to the speech dictionary are used as the noise-reduced representation of noisy speech for feature extraction. A conditional random field (CRF) is then used to model the correlation between feature sequences and voice activity labels along the noisy speech, and voice activity labels are assigned for a given speech observation by decoding the CRF. The presented experimental results in [74] demonstrate that this approach further improves the performance of VAD in low SNR conditions.

Another GMM-based VAD dealing with highly noisy conditions was proposed recently in [75]. In this work,

a VAD has been developed to handle transient noise by using the idea of spectral clustering. Transient noise is a type of sounds that is wrongly detected as speech with high probability, e.g., coughing, sneezing, keyboard typing, and door knocking sounds. Even though there are numerous methods of VAD, this task is challenging in the presence of transient noise. The VAD technique proposed in [75] is a supervised learning algorithm that divides the input speech into two separate clusters, i.e., speech presence and speech absence frames. Labelled data is used in order to adjust the parameters of the kernel in spectral clustering methods for computing the similarity matrix. The parameters obtained in the training stage along with the eigenvectors of the normalized Laplacian of the similarity matrix and the GMM are employed to estimate the likelihood ratio needed for the VAD task. Simulation results prove the high performance of the proposed method, particularly its advantage in treating transient noises, as opposed to the conventional statistical model-based VAD algorithms.

2) Minimum Statistics Tracking

Noise PSD estimation can be accomplished by exploiting the fact that even when speech is present in a time frame, speech energy is not necessarily present in all frequency bins of that time frame. Voiced speech sounds, for example, are quasi-periodic in the time domain and thus are quasi-harmonic in the power spectrum domain, having spectral peaks located at specific frequencies. Therefore, the spectral content between these spectral peaks is representative of the noise PSD. In fact, when a speech spectral peak is present in a frequency bin, the noisy speech power rises far above the noise PSD level. Yet, in many time frames, the noisy power spectrum varies around the true noise PSD level [68]. The minimum statistics (MS) method proposed by Martin [76] mainly uses this observation to estimate the noise PSD without using a VAD. The major idea of this fundamental approach is to collect smoothed noisy periodogram values at each frequency bin, namely $\mathbf{P}(k,l)$, using a sequence of neighboring time frames. By having a large enough number of time frames, i.e., corresponding to around 1 to 2 seconds, it can be guaranteed that the minimum value in the sequence $\mathbf{P}(k,l)$, say $P_{\min}(k,l)$, refers to PSD level without speech presence. However, by considering the minimum of $\mathbf{P}(k,l)$, in general, the distribution of $|V(k,l)|^2$ is sampled below its true mean value, and therefore, $P_{\min}(k,l)$ is an underestimate of $E\{|V(k,l)|^2\}$. In order to fix this issue, a bias compensation is necessarily applied on the minimum statistic, $P_{\min}(k,l)$, as $B_{\min}(k,l).P_{\min}(k,l)$, and the latter term is considered as the noise PSD estimate. However, obtaining this bias is mathematically challenging and approximate bias compensation methods are addressed in [76] and further on in [77].

In practice, although the concept of the MS approach is relatively simple, an efficient implementation of this algorithm with the bias compensation requires a few parameter settings that may not be optimal. Also, a drawback of this MS approach is that by computing the spectral minimum of past time frames, detection of a fast change in the noise level, i.e. in highly non-stationary environments, has an unfavorable amount of delay. Depending on the parameter settings and type of noise, this delay can be as large as one to two seconds. Therefore, abrupt changes in the noise level cannot be generally tracked accurately using the MS approach, resulting in a large amount of residual noise in the underlying noise reduction method due to an underestimated noise PSD. For the aforementioned reasons, there has been a few major modifications and improvements in the literature to the original MS approach. In [78], by making use of a constrained variance smoothing filter, which is actually a generalization of the original MS method in its smoothing parameter, the authors propose a development of the MS method, that is capable of tracking the non-stationary and fast changing behavior of noise more efficiently while reducing its variance. In this approach, the minima of the smoothed periodograms are tracked with a low delay and then they are used to construct VADs, which in turn, are employed to detect the noise-only segments. Finally, the noise PSD is estimated by averaging the noisy periodograms on the noise-only regions.

The original MS method tracks the minimum values of a smoothed power estimate of the noisy speech within a finite search window. To this end, a fixed size for the minimum search window is used regardless of the environmental conditions. However, in [79], the authors suggest to determine variable optimal window lengths according to a variety of noise types. To do this, the window length is selected such that the highest speech quality is achieved depending on individual sources of noise and the underlying speech enhancement technique. The classification of noise in each frame is performed by an ML method which is eventually based on the GMM. As compared to the conventional MS method via various objective and subjective evaluations, it is demonstrated in [79] that the proposed approach provides more accurate noise PSD estimates resulting in the improvement of speech quality.

3) Minimum Controlled Recursive Averaging

The major drawback of the MS method and its variations, however, is that such estimators have a large variance and sometimes they attenuate low energy phonemes.

Another class of soft-decision noise PSD estimators that overcome many drawbacks of the MS method is the so-called improved minima controlled recursive averaging (IMCRA) proposed by Cohen in [80]. This method updates the noise PSD estimate by using a recursive smoothing scheme with time frame and frequency dependent smoothing parameter. The latter is in fact decided by the *a posteriori* speech presence probability (SPP) which is itself controlled by the minima values of a smoothed periodogram. Basically, the estimation of the SPP consists of two iterations of smoothing and minimum tracking. Whereas the first iteration acts as an approximate VAD, the smoothing step in the second iteration excludes relatively strong speech components in the noise PSD estimation. This makes the IMCRA method robust against strong presence of speech components. More specifically, it follows that [80]

$$\begin{aligned}\hat{\sigma}_v^2(k,l) &= \alpha_d(k,l)\,\hat{\sigma}_v^2(k,l-1) + (1-\alpha_d(k,l))|X(k,l)|^2 \\ \alpha_d(k,l) &= \alpha_{d_0} + (1-\alpha_{d_0})p(k,l)\end{aligned} \tag{25}$$

with $p(k,l)$ denoting the *a posteriori* SPP defined as $\mathcal{P}(\mathcal{H}_1|X(k,l))$ and α_{d_0} a fixed smoothing parameter. The latter can be obtained using the noise PSD estimate in the last frame, the *a priori* SNR and the *a priori* speech absence probability, defined as $\mathcal{P}(\mathcal{H}_0)$, with a two-pass searching. The first pass is a coarse estimation where a coarse decision is made to identify the speech and noise components. The second pass is fine searching, which only uses the noise components identified by the first pass searching to calculate the speech presence and absence probabilities. Performance evaluations of the IMCRA approach reported in [80] confirm its noise cancellation advantage over the conventional MS approach when used in a speech enhancement system.

One of the issues with the IMCRA method, however, is that due to the two-pass searching used for local minima tracking, the delay to follow an abrupt noise spectral rise is actually doubled. Also, there still exists considerable speech leakage to the estimated noise PSD by this approach, which causes distortion in the enhanced speech. In [81], an enhanced version of the IMCRA method has been proposed that demonstrates less speech signal leakage and faster response to follow abrupt changes in the noise PSD level. There has also been a few major improvements to the estimation of the *a posteriori* SPP, or in brief SPP, which plays a critical role in the accuracy of the IMCRA method. As such, in [82], an approach to estimate the SPP via HMMs has been proposed. Therein, unlike the conventional SPP which is based solely on the current frame, the temporal correlation present in speech spectra is exploited to obtain the SPP. Specifically, the conventional SPPs are assumed to be the observations of channel-specific two-state HMMs and based on this set of SPPs, the ultimate estimate of the SPPs is obtained via statistical inference techniques such as the forward or forward–backward algorithms. In this sense, the two-state configuration of underlying speech models leads to a low complexity in the HMM processing, and relative to the conventional methods, there is a slight increase in the computational burden.

The SPP in the IMCRA method is commonly calculated independently across the time and frequency in the STFT domain. However, due to the overlap in the STFT frames as well as the correlated nature of the speech signal, there always exist a correlation between subsequent time frames and neighboring frequencies. In this sense some IMCRA-based methods tend to take into account the inherent time and frequency correlations in the calculation of the SPP. In this regard, a major contribution to the IMCRA method has been presented in [83] where the SPP is determined by taking into account the time and frequency correlations in the noisy speech. In this work, by calculating the auto-correlation and cross-correlation across the time and frequency, a primary decision about speech presence is made. Using this decision, the smoothing and weighting parameters in the original IMCRA are refined. Furthermore, the searching process of the local minima is improved by adding a minimum search with a shorter window. Extensive experimental results illustrate that the suggested algorithm improves the accuracy in noise PSD estimation, as compared with the conventional IMCRA. A more recent work that considers inter-frame and inter-band correlations in the STFT domain in the calculation of the SPP has been presented in [84]. Therein, it has been shown that the detection accuracy of the SPP estimators can be increased by taking into account only a few neighboring time frames and frequency bins.

The conventional Bayesian approach to estimate the SPP is based on a likelihood ratio that is derived by assuming a Gaussian distribution for the speech. However, some recent developments on the SPP estimation consider likelihoods of speech presence based on super-Gaussian speech models or, alternatively, based on averaged observations. As such, in [85], these two aspects are combined and a closed-form solution for the generalized likelihood of speech presence has been derived. Furthermore, contrary to the conventional IMCRA methods, in order to obtain SPP estimates close to zero in the speech absence, fixed values for the *a priori* SNR and *a priori* SPP have been employed. The proposed improved SPP estimation is shown to outperform SPP estimation methods that consider averaged observations with a Gaussian speech model, a super-Gaussian model with no averaging, and also the conventional IMCRA.

D. Performance Evaluation of Speech Enhancement

In order to judge the efficiency of a speech enhancement system, the quality of the enhanced speech at its output has to be evaluated. In principle, there are two types of methods for the evaluation of speech quality: subjective and objective methods. The former is based on listening tests performed by human and the latter is done by calculating the so-called objective performance measures. Even though the most accurate and reliable way for evaluating speech quality is to perform subjective listening tests, they are often costly and time consuming given that they have to be performed under stringiest conditions [86]. For this reason, much effort has been made to develop objective measures that would be able to determine speech quality consistently with subjective tests. The most important objective performance measures include segmental SNR (segSNR), weighted-slope spectral (WSS) distance, perceptual evaluation of speech quality (PESQ) and the linear prediction coefficients (LPC)-based objective measures [86]. In this part, we briefly discuss a handful of the most commonly used objective and subjective measures for speech quality.

The basic version of the segSNR measure in the time domain can be expressed as [87]

$$\text{segSNR} = \frac{1}{L}\sum_{l=1}^{L} 10\log_{10}\left(\frac{\|\mathbf{s}_l\|^2}{\|\mathbf{s}_l - \hat{\mathbf{s}}_l\|^2}\right) \tag{26}$$

where $\mathbf{s}_l$ and $\hat{\mathbf{s}}_l$ are respectively the lth frame of the clean and enhanced speech, L is the number of total frames and $\|\cdot\|$ denotes the Euclidean norm. To avoid reaching unreasonably high or low values for the segSNR, high and low thresholds are often set on this quantity. The segSNR can also be defined in the frequency (STFT) domain in which case it is usually referred to as the frequency weighted segSNR. The proper weighting can be taken as a raised power of the amplitude spectrum of the clean speech in order to emphasize on the time-frequency units where clean speech exists [86].

The WSS distance measure is a direct spectral distance measure that is based on the comparison between the smoothed spectra from the clean and distorted speech. Since the smoothed spectra can be obtained by different approaches such as using linear prediction (LP) analysis, cepstrum liftering (a term coined for filtering in the cepstrum domain), or a filter bank analysis, it can be implemented in different ways. One famous expression for this measure is as follows [88]

$$\text{d}_{\text{WSS}} = \frac{1}{L}\sum_{l=1}^{L}\frac{\sum_{k=1}^{K} W(k,l)(S_c(k,l) - S_d(k,l))^2}{\sum_{k=1}^{K} W(k,l)} \tag{27}$$

where $S_c(k,l)$ and $S_d(k,l)$ are the spectral slopes typically defined as the spectral differences between neighboring frequency bins and $W(k,l)$ is a proper weighting.

The PESQ measure is one of the most computationally complex objective measures and is the one recommended by the International Telecommunication Union-Telecommunication Standardization Sector (ITU-T) for speech quality assessment of handset telephony and narrow-band speech codecs [89]. Basically, the PESQ score is calculated as a linear combination of the average disturbance value D_{ind} and the average asymmetrical disturbance value A_{ind} as follows

$$\text{PESQ} = a_0 + a_1 D_{\text{ind}} + a_2 A_{\text{ind}} \tag{28}$$

Note that the primarily suggested values for the three parameters a_0, a_1 and a_2 are optimized for speech processed through networks and not through noise reduction methods. However, in [86], three modifications of this parameter set have been suggested that make the PESQ measure suitable for evaluating speech distortion, noise distortion and overall speech quality.

Another popular group of objective quality measures is the LPC-based objective scores which include the log-likelihood ratio (LLR), the Itakura-Saito (IS), and the cepstrum distance measures as the most important [86]. The LLR measure is defined as

$$\text{d}_{\text{LLR}} = \log\left(\frac{\mathbf{a}_p \mathbf{R}_c \mathbf{a}_p^T}{\mathbf{a}_c \mathbf{R}_c \mathbf{a}_c^T}\right) \tag{29}$$

where $\mathbf{a}_c$ and $\mathbf{a}_p$ are respectively the LPC coefficient vectors of the clean and enhanced speech frame and $\mathbf{R}_c$ is the autocorrelation matrix of $\mathbf{a}_c$. In practice, the smallest 95% of the frame LLR values may be used to calculate the average LLR and the frame LLR values are also limited to the range of $[0,2]$. The IS measure is defined as

$$\text{d}_{\text{IS}} = \frac{\sigma_c^2}{\sigma_p^2}\left(\frac{\mathbf{a}_p \mathbf{R}_c \mathbf{a}_p^T}{\mathbf{a}_c \mathbf{R}_c \mathbf{a}_c^T}\right) + \log\left(\frac{\sigma_c^2}{\sigma_p^2}\right) - 1 \tag{30}$$

where σ_c^2 and σ_p^2 are respectively the LPC gains of the clean and enhanced speech. The IS values are typically limited in the range of $[0,100]$. An objective measure based on cepstrum coefficients can be computed as the following

$$\text{d}_{\text{CEP}} = \frac{10}{\log 10}\sqrt{2\sum_{k=1}^{p}(c_c(k) - c_p(k))^2} \tag{31}$$

where $c_c(k)$ and $c_p(k)$ are the cepstrum coefficients of the clean and enhanced speech, respectively, and p is the order of the underlying LPC analysis. The cepstrum coefficients can be in turn calculated from the LPC coefficients, $a(k)$, in a recursive manner, using the following expression

$$c(k) = a(k) + \sum_{k'=1}^{k-1} \frac{k'}{k} c(k')\, a(k-k'), \quad 1 \le k \le p \qquad (32)$$

To minimize the number of outliers, the cepstrum distance can be limited to the range of [0], [10].

Since the aforementioned objective performance measures cannot be relied to be fully correlated with speech/noise distortions and overall speech quality, composite objective measures may come useful [86]. These measures can be obtained through combining the introduced basic objective measures by utilizing techniques such as multiple linear regression analysis, e.g. [90].

Subjective quality measures are based on the subjective opinion of a group of listeners on the quality of enhanced speech samples. The most famous subjective quality measures for speech transmission over voice communication systems have been recognized and standardized by ITU-T. In this regard, opinion rating methods can be used to evaluate the overall perception quality of a speech sample. The mean opinion score (MOS), initially developed for telephone bandwidth speech, is one of the most widely used opinion rating methods. In MOS, listeners are required to rate the speech sample under the test into one of the five quality categories. Each category is represented by a number from 5 to 1, and it corresponds to an excellent, good, fair, poor and unsatisfactory speech quality. The ultimate MOS value is the average of all listeners for each of the speech samples under the test. Clearly, the enhanced speech in general suffers from various aspects of degradation including bandwidth limitation, additive noise, echo and nonlinear distortions. The MOS measure provides an overall impression of all different degradations, measured as one numerical value [88].

III. Recent Advances

In this section, we discuss the major aspects in the recent development of STFT-based noise reduction methods for single- and multi-channel cases. With a focus on the most recent literature, we elaborate in the single-channel case, on the spectral subtractive method in the modulation domain, STSA estimation using non-Gaussian speech prior models, estimation of the two major noise reduction parameters, i.e. the noise PSD and the *a priori* SNR, and the estimation of speech spectral phase. In the multi-channel case, we investigate the extension of STSA estimators from single to multiple channel, the estimation of noise PSD matrix (mostly used in the MVDR beamformer) and some recent advances in the multi-channel Wiener filtering.

A. Use of Super-Gaussian Speech Priors

The speech STSA estimators discussed in Section II and the MMSE amplitude estimators represented in Table 2 are all based on the Rayleigh distribution for speech STSA. The latter arises from the fact that speech STFT coefficients are generally assumed to have a complex Gaussian distribution. Recently, however, there has been numerous works directed towards the estimation of speech STSA using super-Gaussian statistical models, especially for the speech STSA. In [91] and references therein, various non-Gaussian distribution models for the speech STSA are discussed, which include exponential, Laplacian, Chi, Gamma (one-sided) and generalized Gamma distributions. These distributions each have unknown parameters and different speech data-based (adaptive) schemes have been proposed for the estimation of their corresponding parameters. According to the experiments in [92], [93], the generalized Gamma distribution (GGD) has the potential to fit the empirical (e.g. histogram-based) distribution of speech STSAs best, however, closed-form solutions for an STSA estimator is available only for specific choices of the parameters of the GGD. In fact, the GGD is a very flexible parametric distribution which covers many super-Gaussian distributions as particular cases. The one-sided GGD family with shape parameters a and c and scaling parameter b is given by [94]

$$p_{\mathrm{GGD}}(A;a,b,c) = \frac{ab^c}{\Gamma(c)} A^{ac-1} \exp(-bA^a);\ A \ge 0,\ a,b,c > 0 \qquad (33)$$

with a and c as the shape parameters and b as the scaling parameter. Note that since this subsection deals with spectral amplitude estimation, only right-sided distributions are discussed. In fact, the GGD model is a very generalized form of different super-Gaussian distributions and a few useful super-Gaussian distributions in the context of STSA estimation can be derived by considering particular choices of the GGD model, which is summarized in Table 4.

Fig. 4 shows GGD values for a few choices of its shaping parameters and $b = 2$. This indicates that by a dynamic selection of these parameters at each STFT frequency bin and time frame, one can gain control over the statistical model of the speech STSA and thus the corresponding gain function of STSA estimators. In a theoretical viewpoint, the estimation of GGD

Table 4.
Parameter sets of the GGD leading to Rayleigh, Gamma, Chi, or exponential speech spectral amplitude models.

Parameters of the GGD	STSA Prior
$a = 2, c = 1$	Rayleigh
$a = 1$	Gamma
$a = 2, b = 1/2$	Chi
$a = 1, c = 1$	Exponential

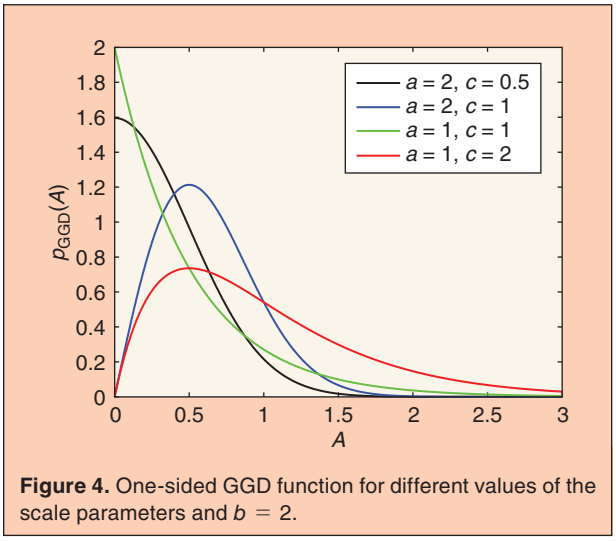

Figure 4. One-sided GGD function for different values of the scale parameters and $b = 2$.

parameters can be done through an ML procedure using the available noisy speech data. However, the exact determination of the GGD parameters independently by solving likelihood equations is cumbersome [93]. In the context of speech STSA estimation, however, closed-form solutions (for ML, MAP or MMSE-based) estimators are available only for the choices of $a = 1$ and $a = 2$. Note that for the choice of $a = 2$, the GGD prior is actually simplified into a generalized form of the Chi distribution with 2c degrees of freedom and $1/\sqrt{2b}$ as the scale parameter. Also, the second moment of the GGD prior in (33), i.e. the speech STSA variance, is given as the following [94]

$$\sigma_A^2 = \begin{cases} \dfrac{c(c+1)}{b^2}, & \text{if } a = 1 \\ \dfrac{c}{b}, & \text{if } a = 2 \end{cases} \tag{34}$$

Therefore, having an estimation of the speech STSA spectral variance, σ_A^2, the scale parameter b will be obtained based on the choice of the shaping parameters. Various combinations of the GGD shape parameters that lead to specific closed-form solutions for speech STSA estimators have been presented in [94]. Therein, solutions have been presented for the case of MMSE-based estimators using Gaussian and exponential speech priors, and MAP estimators using GGD speech priors with $a = 1, 2$. It is concluded that in the case of MMSE-based estimation, higher order shape parameters generally results in numerical analysis since such expressions rely on integrations with no closed-form solution. Also, in the case of MAP estimators, certain combinations of lower order shape parameters can result in monotonic cost functions for which a MAP solution does not actually exist. STSA estimation solutions using special cases of the GGD for noise distribution have also been discussed in [94], yet, in accordance with the results reported in [13], no improvements have been obtained as compared to using the Gaussian distribution for noise. Table 5 summarizes the major solutions of STSA estimation using the GGD prior presented in [94]. Note that in this table, (a_s, c_s) and (a_v, c_v) respectively denote the GGD shape parameters for the clean speech and noise priors.

In [95], a family of log-spectral amplitude (LSA) estimators have been proposed, using GGD priors with $a = 1, 2$. Therein, due to providing mathematical flexibility in the statistical STSA modelling, objective improvements with respect to several older STSA estimators including the LSA estimator in [20] have been achieved. Although closed-form solutions are not obtainable for the general case of $a = 1, 2$, estimators were expressed in [95] as limits, and were mathematically approximated. In [96], MMSE-based and MAP estimators of speech STSA have been proposed based on Gamma and Chi priors for speech STSA, and data-driven schemes for the selection of the shape parameter of the priors have been suggested. In that work, rather than relying on *a priori* estimated values of the shape parameter, the focus is on seeking those values that maximize the quality of the enhanced speech, in an *a posteriori* fashion. To this end, the performance of the parameter selection schemes is first evaluated as a function of the shape parameter and then optimal values are found by means of a formal subjective listening test. The main conclusion was that the shape parameters control a trade off between the level of the residual noise and its musical character. Also, it was found that the optimal parameter values maximizing the subjective performance are different than those maximizing the scores in objective performance measures. It is believed that this discrepancy is mainly due to the poor ability of objective measures to penalize the musical noise artifacts. Another finding of the research in [96] is that very close performance results can be obtained using the same estimator, i.e. MMSE-based or MAP, but with different STSA priors. This can be attributed to the flexibility provided by the shape parameters of the STSA prior, allowing the listener to closely match the performance of two estimators with different speech priors. As further conclusions of this work, the type of the estimators, i.e. MMSE-based or MAP, has significant impact on the quality of the enhanced speech. Whereas MAP estimators result in lower residual noise levels, the MMSE-based estimators are more successful in the restoration of the speech spectral components and are able to achieve higher scores in the objective speech quality measures. Both type of STSA estimators, however, can produce an enhanced speech free of musical noise artifacts, given the correct setting of their parameters.

Table 5.
Speech STSA estimators for particular parameter choices of the GGD speech and noise priors.

Criterion	(a_v, c_v)	(a_s, c_s)	Gain Function
ML	$(a_v \in \{1,2\}, c_v \geq 1/a_v)$	–	$1-\frac{1}{\sqrt{\gamma}}\left(\frac{a_v c_v - 1}{a_v\sqrt{c_v(c_v+2-a_v)}}\right)^{1/a_v}$
MMSE	$(2, 1/2)$	$(2, 1/2)$	$\frac{\zeta}{1+\zeta}\left[1-\frac{1}{\gamma}\exp\left(-\frac{\gamma(1+\zeta^2)}{4\zeta(1+\zeta)}\right)\sinh\left(\frac{\gamma(\zeta-1)}{4\zeta}\right)\right]$
		$(1, 1)$	$1-\frac{1}{2\sqrt{\zeta\gamma}}-\frac{1}{\gamma}\left[\exp\left(-\frac{1}{\zeta}-\frac{\gamma}{4}+\sqrt{\frac{\gamma}{2\zeta}}\right)\sinh\left(\frac{\gamma}{4}-\sqrt{\frac{\gamma}{2\zeta}}\right)\right]$
	$(1, 1)$	$(1, c_s \in \mathbb{N})$	$\frac{1}{\mathcal{A}}\left\{\frac{(-1)^{c_s}c_s! + \exp(\mathcal{A})\sum_{n=0}^{c_s}\left((-1)^n\frac{c_s!}{(c_s-n)!}\mathcal{A}^{c_s-n}\right)}{(-1)^{c_s-1}(c_s-1)! + \exp(\mathcal{A})\sum_{n=0}^{c_s-1}\left((-1)^n\frac{(c_s-1)!}{(c_s-n-1)!}\mathcal{A}^{c_s-n-1}\right)}\right\}$ where $\mathcal{A}=\sqrt{\frac{2\gamma}{\zeta}}\left(\sqrt{\zeta}-\sqrt{\frac{c_s(c_s-1)}{2}}\right)$
MAP	$(2, 1/2)$	$(2, c_s \in \mathbb{R})$	$\frac{1}{2(2c_s+\zeta)}\left(\zeta+\sqrt{\zeta^2+4(\zeta/\gamma)(2c_s-1)(2c_s+\zeta)}\right)$
		$(1, c_s \in \mathbb{R})$	$\frac{1}{2}-\frac{\sqrt{c_s(c_s+1)}}{2\sqrt{\gamma\zeta}}+\sqrt{\left(\frac{1}{2}-\frac{\sqrt{c_s(c_s+1)}}{2\sqrt{\gamma\zeta}}\right)^2+\frac{c_s-1}{\gamma}}$
	$(2, 1)$	$(2, 1/2)$	$\frac{1}{2}\left(\frac{4\zeta+1}{2\zeta+1}\right)+\frac{1}{2}\sqrt{\left(\frac{4\zeta+1}{2\zeta+1}\right)^2-\frac{4\zeta}{\gamma}\left(\frac{2\gamma-1}{2\zeta+1}\right)}$
		$(1, 1)$	$1-\frac{1}{2\sqrt{2\zeta\gamma}}+\sqrt{\left(1-\frac{1}{2\sqrt{2\zeta\gamma}}\right)^2+\left(\frac{1}{\sqrt{2\zeta\gamma}}+\frac{1}{2\gamma}-1\right)}$

In [97], a generalized MAP estimator using the Gamma STSA prior along with a data-driven scheme to estimate its shape parameter has been proposed. The shape parameter scheme is based on the fact that a higher estimated SNR corresponds to stronger presence of speech components with respect to noise, and thus, a higher gain value is required for speech segments with higher SNRs. Therefore, since the derived gain function is monotonically decreasing with the Gamma shape parameter, the proposed parameter scheme suggests lower shape parameters for higher SNRs and vice versa. Performance comparisons with other conventional STSA estimators, i.e. the MMSE, ML and MAP methods, confirms that the suggested MAP estimator provides better objective scores in low SNRs while having comparable performance in high SNRs.

The STSA estimators discussed so far incorporated improved statistical models with the original MMSE or log-MMSE cost functions. In [98], the authors make use of the Chi STSA prior to derive estimators using perceptually motivated spectral amplitude cost functions, namely the WE and WCOSH primarily developed in [21]. The major purpose in [98] is to determine the advantage of incorporating improved cost functions with more accurate (i.e. super-Gaussian) STSA priors. Therein, it was shown that whereas the perceptually-motivated cost functions emphasize spectral valleys rather than spectral peaks (formants) and indirectly account for auditory masking effects, the incorporation of the Chi STSA prior demonstrates considerable improvement over the Rayleigh model for the speech prior. Yet, no systematic parameter choice has been proposed for the two WE and WCOSH estimators and the shape parameter of the corresponding Chi STSA prior is selected empirically. Along the same line of work, in [99], the authors take advantages of the β-order MMSE cost function firstly adopted in [22] with Laplacian priors for the real and imaginary parts of speech STFT coefficients. Even though using Laplacian model as speech prior primarily results in a highly non-linear estimator with no closed-form solution and high computation costs, by using approximations for the distribution of speech STFT and also for the involved Bessel functions, an improved closed-form version of the estimator has been derived and evaluated in [99]. The comparative evaluations confirm the superiority of the suggested estimator relative to the state-of-the-art estimators that assume either Gaussian or Laplacian STSA priors such as [100]. Finally in [101], a general form of an STSA estimator under Wβ-SA cost function and

GGD speech prior has been derived. New schemes have then been proposed for the estimation of the cost function as well as speech prior parameters using: an initial estimate of the speech STSA, the noise masking feature of the human auditory system and the estimated SNR. It is concluded that the exploitation of a primary STSA estimate in the parameter selection for speech prior leads to more efficient control on the gain function values. Objective performance evaluation in different noise conditions demonstrates the superiority of the proposed estimator over the state-of-the-art STSA estimators.

Based on the aforementioned works, it can be concluded that even though statistical methods for the estimation of the parameters of super-Gaussian priors exist, e.g. [102], subjectively driven schemes based on speech observations or solid theoretical methods to maximize objective measures, such as [96], [97], prove to be more efficient.

B. Advances in Noise PSD Estimation

In Section II, three main methods of noise PSD estimation and their improvements were studied. However, recently there has been growing interest in the statistical model-based estimation of noise PSD, which also takes advantage of some of the concepts in the older methods. These statistical model-based methods allow to relax the assumption that time-frequency units can be found where only the noise component is dominant. Note that the latter assumption along with the assumption that noise is generally more stationary than speech are the basis for noise estimation methods like MS and IMCRA. Using statistical models for the noise PSD, however, enables the estimation of noise PSD even if speech signal is dominant in a time-frequency unit. We present in this part the most important of these statistical methods, including MAP, MMSE and ML noise estimation.

1) MAP Estimation

In [103], [104], the authors address the MAP estimation of noise PSD for a general non-stationary noisy environment. This is made possible only by assuming the availability of an initial estimate of the speech PSD. Therefore, the suggested MAP-based algorithm is meant to be used as a post-processor to a first speech enhancement stage. The estimation of the parameters of the noise process is then reduced into the problem of estimating the variance of a complex-valued zero-mean white Gaussian random process using noisy observations, which is solved by a MAP-based estimation method in [103]. The major advantage with this approach is the ability to follow non-stationary noise dominated by strong speech components even in the critical case of rapidly rising noise level. The presented experimental comparison with the state-of-the-art noise tracking algorithms demonstrates smaller estimation errors under low SNR conditions and smaller fluctuations of the estimated noise PSD values. In [104], an improved version of the MAP-based noise estimation has been proposed, where an empirical bias compensation and bandwidth adjustment are suggested to reduce the bias and variance of the noise PSD estimate, resulting in smaller estimation variance.

2) MMSE Estimation

In [105], an MMSE estimator of the noise PSD with low complexity has been firstly proposed that is highly useful for applications with low-complexity constraints such as hearing aids. Therein, an MMSE estimator of the noise amplitude-squared STFT coefficients has been formulated in the Bayesian framework, as the following [105]

$$\hat{\sigma}_v^2 = E\{\mathcal{V}^2 \,|\, X\} = \int_0^{+\infty} \mathcal{V}^2 \, p_{\mathcal{V}|X}(\mathcal{V})\, d\mathcal{V} \tag{35}$$

where $\mathcal{V}$ denotes the noise amplitude and $p_{\mathcal{V}|X}(\mathcal{V})$ is its distribution conditioned on the speech observation. Assuming that both the speech and noise STFT coefficients are modelled by a complex Gaussian distribution, it follows that

$$\hat{\sigma}_v^2 = \left(\frac{1}{(1+\zeta^2)} + \frac{\zeta}{(1+\zeta)\gamma} \right) |X|^2 \tag{36}$$

giving the noise PSD estimate as an instantaneous function of the *a priori* and *a posteriori* SNRs. Even though this estimator is generally unbiased, due to the inaccuracies in the estimation of the *a priori* SNR, a necessary bias compensation is suggested in [105]. When used in a speech enhancement system, this MMSE-based noise tracking provides superior performance compared to the MS-based method. Yet, compared with the state-of-the-art noise PSD estimation methods, the MMSE-based approach has almost similar performance but with a considerably lower computational burden. In [106], the authors further analyze the MMSE-based approach and suggest to use it in a recursive smoothing scheme. Next, by suggesting a modified version of the original MMSE-based noise estimation, its performance is further improved. In this sense, the original method is firstly interpreted as a VAD-based noise PSD estimator, and next, it is shown that the bias compensation step is unnecessary if the VAD is replaced by a soft-decision SPP-based method with fixed priors. The use of SPP with fixed optimally set priors is advantageous in this approach, as compared to the IMCRA method where the priors are adaptively determined at each time-frequency unit. In [106], the following expression for the SPP is derived

$$\mathcal{P}(\mathcal{H}_1 \,|\, X) = \left(1 + \frac{\mathcal{P}(\mathcal{H}_0)}{\mathcal{P}(\mathcal{H}_1)} (1+\zeta_{\mathcal{H}_1}) e^{-\frac{|X|^2}{\hat{\sigma}_v^2} \frac{\zeta_{\mathcal{H}_1}}{1+\zeta_{\mathcal{H}_1}}} \right)^{-1} \tag{37}$$

with fixed *a priori* SPPs assumed as $\mathcal{P}(\mathcal{H}_1) = \mathcal{P}(\mathcal{H}_1), \hat{\sigma}_v^2$ evaluated by the estimated PSD at the previous frame and $\zeta_{\mathcal{H}_1}$ denoting the optimal value for the *a priori* SNR under $\mathcal{H}_1$. The latter is obtained by minimizing the total error probability in the speech presence/absence hypothesis testing [106]. In terms of performance, it is demonstrated that the proposed MMSE- and SPP-based approach maintains the quick noise tracking performance of the original MMSE-based method while exhibiting less overestimation of the noise PSD and having an even lower computational complexity. In [107], the same authors further work on the MMSE-based noise estimator by employing a more advanced estimator of the speech PSD, instead of the conventional ML-based *a priori* SNR estimator, based on temporal cepstrum smoothing (TCS). The latter, even though implies heavier computational burden, is able to provide a more precise estimate of the speech PSD through exploiting knowledge about the speech spectral structure. Moreover, the requirement for the bias compensation in the original approach, that is fulfilled by using the decision-directed method to estimate the speech PSD, is eliminated. Using this noise PSD estimator in a noise reduction framework, it is concluded that a higher noise reduction performance with a comparable amount of speech distortion is achieved.

3) ML Estimation

Another major method in the category of statistical model-based noise PSD estimators has been proposed in [108] wherein a recursive ML-based noise estimation algorithm is derived. Within this approach, the *a priori* and *a posteriori* SPPs as well as noise statistics are analytically retrieved from an expectation maximization (EM) algorithm at every time-frequency unit. The recursive updating of these three terms are performed in a unified manner through new closed-form expressions and without relying on the conventional tracking of speech PSD minima. As compared with the MS, IMCRA and MMSE-based approaches, the aforementioned approach is optimal in the ML sense and requires only one tuning parameter, i.e. the forgetting factor in the corresponding recursive smoothing process. In addition to its low computational load, the ML-based noise estimator can achieve a performance level superior or comparable to that of the previous algorithms.

4) Evaluation of Noise PSD Estimation Methods

The diversity of the proposed methods for the estimation of noise PSD and the growing interest in this field makes it necessary to compare the most well-known approaches in a unified framework. In [109], the authors investigate the performance of several major recent approaches of noise PSD estimation and some of their variations in adverse acoustic environments. In this evaluation, to be independent of an underlying noise reduction method, the standalone performance of the noise PSD estimators is measured with respect to the reference noise PSD that is obtained by smoothing the noise STFT. To do so, both the mean of a spectral distance measure and the variance of the estimators are comparatively assessed and through a variety of non-stationary noise types, the robustness of the noise estimators in adverse environments is examined. In this regard, a total of 8 algorithms with their variations including the MS [76] method, the IMCRA [80] method, a subspace decomposition based approach [110] and the MMSE-based method [105] have been implemented and compared.

First of all, it is concluded that any deviation from the specific parameter setting suggested by each algorithm, such as that for the amount of STFT overlap, deteriorates the performance of the algorithm under investigation. Next, it is found that some of the noise estimators are more sensitive to the level of SNR than the others. In this sense, while the IMCRA method performs very well in low SNRs, by increasing the SNR, the estimation error measure increases for this method. This is despite the fact that the MS and MMSE-based methods are more robust to the fluctuations in the noise level. Another main conclusion is that when the noise PSD does not change rapidly in the time, most of the algorithms perform similarly, whereas for non-stationary noise a few methods show to be more robust. The most robust noise estimator, however, based on the experiments in [109], is found to be the MMSE-based method in [105]. The subspace decomposition based approach in [110] also provides similar performance in most of the noise scenarios with inferior performance in others.

At the end of this section, it should be noted that even though there are numerous methods in the literature targeting the noise PSD estimation topic, this problem is still challenging in low SNR conditions and fast-changing non-stationary noise environments. Therefore, there is room for further research on this topic in the future in order to improve the accuracy (reduce the error variance) of the current approaches and make their tracking ability robust against sudden changes in the noise level.

C. Speech Enhancement in the Modulation Domain

Modulation domain processing has found applications in areas such as speech coding, speech recognition, speaker recognition, objective speech intelligibility evaluation as well as speech enhancement. There is considerable psychoacoustic and physiological evidence to support the importance of the modulation domain in speech signal processing. While the envelope of the acoustic

amplitude spectrum is capable to represent the shape of the vocal tract, the modulation spectrum represents how the vocal tract changes by time. In fact, it is these temporal changes that convey most of the linguistic information (or intelligibility) of speech signal [111].

The speech enhancement techniques discussed so far employ the analysis-modification-synthesis (AMS) framework to perform enhancement in the acoustic spectral domain. Speech enhancement in the modulation spectral domain is an extension of the acoustic AMS framework to include modulation domain processing features. Firstly introduced by Atlas et al. [112], the short-time modulation spectrum is basically a function of time, acoustic frequency and modulation frequency. Whereas the acoustic spectrum is the STFT of the speech signal, the modulation spectrum at a given acoustic frequency is the STFT of the time series corresponding to the acoustic spectral amplitudes at that frequency. More specifically, the modulation spectrum can be expressed using STFT analysis as the following [113]

$$\mathcal{X}(\eta,k,m)=\sum_{l=-\infty}^{\infty}|X(k,l)|w(\eta-l)e^{-j2\pi ml/M} \tag{38}$$

where η is the acoustic frame number, m refers to the index of the discrete modulation frequency, M is the modulation frame duration (in terms of acoustic frames) and $w(\eta)$ denotes a modulation analysis window function. The spectral subtractive methods discussed in Section II can now be implemented in the short-time modulation domain defined by (38) instead of the conventional STFT domain. In [113], to obtain the best performance, it is proposed to use the following generalized spectral subtraction in the modulation domain

$$\hat{\mathcal{A}}(\eta,k,m)=\begin{cases}(|\mathcal{X}(\eta,k,m)|^{\rho_1}-\rho_2|\hat{\mathcal{D}}(\eta,k,m)|^{\rho_1})^{1/\rho_1}, \\ \quad if\ |\mathcal{X}(\eta,k,m)|^{\rho_1}-\rho_2|\hat{\mathcal{D}}(\eta,k,m)|^{\rho_1} \\ \quad\quad \geq \rho_3|\hat{\mathcal{D}}(\eta,k,m)|^{\rho_1} \\ (\rho_3|\hat{\mathcal{D}}(\eta,k,m)|^{\rho_1})^{1/\rho_1}, \quad \text{otherwise}\end{cases} \tag{39}$$

with ρ_1, ρ_2 and ρ_3 the fix parameters of the spectral subtraction scheme. The estimate of the modulation amplitude spectrum of the noise, $\hat{\mathcal{D}}(\eta,k,m)$, is obtained using a decision from a simple voice activity detector (VAD) applied in the modulation domain. Note that, in order to keep the estimated modulation amplitude spectrum, $\hat{\mathcal{A}}(\eta,k,m)$, from approaching very small values, a flooring scheme has been applied in the second branch of (39). It is stated in [113] that unlike the acoustic phase spectrum, the modulation phase spectrum contains important information about the speech signal, yet, the estimated modulation phase spectrum is taken as the acoustic phase spectrum of noisy speech. Finally, the estimate of the modified acoustic amplitude spectrum, i.e. the STFT amplitude $\hat{A}(k,l)$, is obtained by taking the inverse STFT of $\hat{\mathcal{A}}(\eta,k,m)e^{j\Theta_X(\eta,k,m)}$ followed by overlap-add with synthesis windowing. A block diagram of the general scheme of the spectral subtraction in the modulation domain is shown in Fig. 5. The most challenging problem with the speech enhancement in the modulation domain, however, is that the enhanced speech generally suffers from temporal slurring distortion which appears as speech unintelligibility. On the other hand, the conventional STFT domain technique does not suffer from the slurring distortion, even

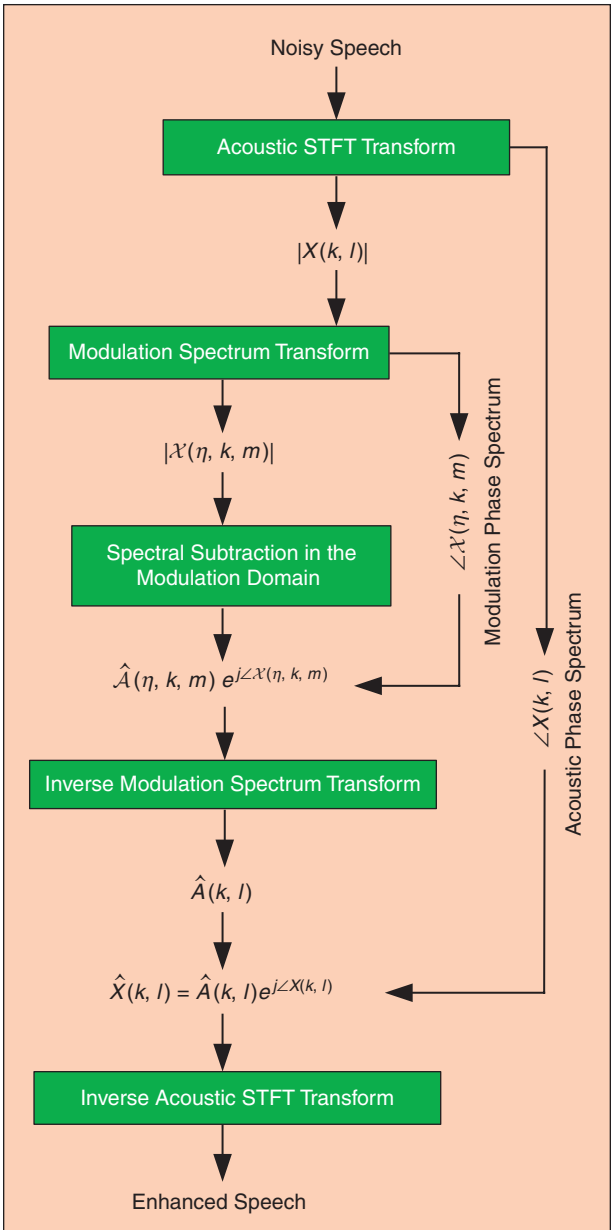

Figure 5. Block diagram of the modulation domain speech enhancement procedure [113].

though it is less effective at removal of background noise and suffers from the musical noise effect. Thus, it was proposed in [113] to exploit the strengths of the two methods, while trying to avoid their weaknesses, by combining (fusing) them in the acoustic STFT domain, as the following scheme

$$\hat{A}(k,l) = \Phi(\gamma)\,\hat{A}^{(1)}(k,l) + (1-\Phi(\gamma))\,\hat{A}^{(2)}(k,l) \qquad (40)$$

with $\hat{A}^{(1)}(k,l)$ and $\hat{A}^{(2)}(k,l)$ denoting the STFT amplitude spectra estimated through the acoustic and modulation STFT techniques, respectively, and $\Phi(\gamma)$ is the fusion weighting function. The latter is suggested to be an increasing function of the *a posteriori* SNR γ, in order to give emphasis to the modulation domain estimation in lower SNRs while favoring the acoustic domain estimate in higher SNRs. Using an objective speech quality measure, namely the perceptual evaluation of speech quality (PESQ) as well as formal subjective listening tests, it was shown in [113] that the MMSE-based modulation domain technique results in improved speech quality with respect to the MMSE-based acoustic STFT domain technique. Furthermore, the fusion of the two techniques, i.e. the modulation and the acoustic domain, as in (40), results in further improvements due to the good compromise between different types of spectral distortions, namely musical noise and temporal slurring.

In [114], a few contributions to the basic MMSE-based modulation domain method in [113] have been made and investigated. These include the extension from the MMSE cost function to the log-MMSE case and use of speech presence probability (SPP) to involve the uncertainty in the presence of speech in the gain function. In [115], motivated by psychoacoustic evidence of frequency selectivity in the modulation domain, the authors introduce the concept of frequency channel selection in the spectral modulation domain as a potential means of improving the speech intelligibility. Therein, the SNR measure in the spectral modulation domain is employed to identify the modulation frequencies dominated by speech and those dominated by noise. Next, the speech-dominated modulation frequencies are retained, whereas the noise-dominated ones, which are detrimental to speech intelligibility, are discarded. This work has further been shown to be beneficial in speech recognition in the presence of noise. In [116], the famous method of Kalman filtering is investigated in the modulation spectrum domain and its performance evaluation against time domain and STFT domain Kalman filtering demonstrates superiority in noise reduction and minimal speech distortion. This is because the Kalman filter is basically a joint amplitude and phase spectrum estimator, it is highly useful in modulation-domain processing, as phase information plays an important role in the modulation spectrum domain contrary to the acoustic spectrum domain [113]. Also in [117], in order to preserve the phase information of speech in the modulation spectrum domain, it is suggested to perform the spectral enhancement on the real and imaginary parts of the complex-valued spectra separately. Objective and subjective evaluation experiments indicate that the proposed method outperforms the modulation domain amplitude spectral subtraction, nonlinear spectral subtraction and the conventional MMSE-based estimator in the STFT domain.

D. Speech Phase Estimation

Typical speech enhancement methods in the STFT domain modify only the amplitude spectrum and keep the phase spectrum of noisy speech unchanged. However, especially recently, there has been growing interest in the estimation of the speech spectral phase and aiming at further improvement of speech quality by employing a closer estimate of speech phase to the clean one than the noisy phase. In this subsection, we explain the reason why most of the research in STFT-based noise reduction has been focused on amplitude estimation and why spectral phase estimation is recently becoming more attractive in the literature. Next, we present a brief overview of the most major and recent spectral phase estimation approaches in the STFT domain and draw a few conclusions. To this end, we mainly take advantage of the work by T. Gerkmann et al in [118] which presents a comprehensive overview on the history and advances in speech phase processing.

Early in the literature of speech enhancement in the frequency domain, Wang and Lim [14] conducted experiments where the speech spectral phase was corrupted by white noise at different SNRs, while the amplitude remained the same. Their conclusion was that the difference in the output SNR for speech signals with various noisy phases is not considerable, and therefore, enhancement of speech phase is not really important. Also, Ephraim and Malah [16] proved theoretically that, given the independence of speech spectral amplitude and phase, the MMSE optimal estimate of the spectral phase is actually the noise phase. Note that the independence of speech spectral amplitude and phase is not generally true but is assumed to make the speech amplitude estimation mathematically tractable. As a consequence, the focus of attention in the spectral enhancement techniques has mainly been on improving the speech amplitude with using the unchanged noisy phase.

However, with the increase in processing power, researchers have started investigating the role of spectral

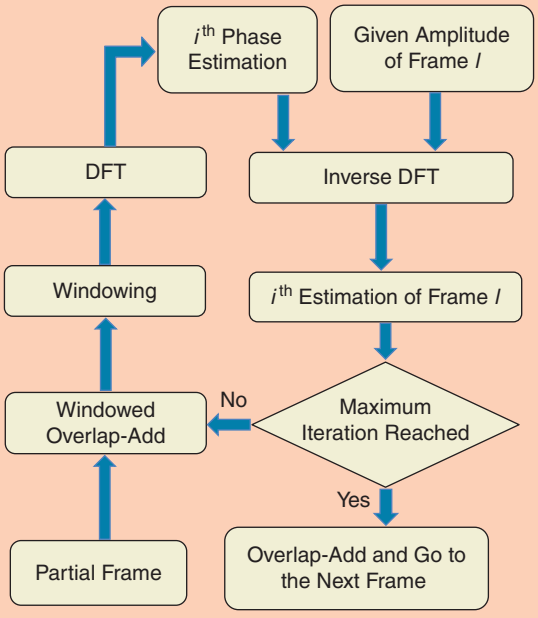

Figure 6. Block diagram representation of the basic version of the RTISI method for phase estimation [120].

phase in improving the speech quality over the past few years. Paliwal et al. [119] demonstrated through extensive subjective and objective performance evaluations that, given the STFT overlap is increased a bit, the performance of amplitude estimators can be significantly improved if combined with less noisy spectral phases. Also in [118], experiments have been done to show that part of the speech signal's structure can be drawn from its phase in the STFT domain. Therein, spectrogram inspections have been done for speech phase, group delay and instantaneous frequency (IF). It is shown that, even though the spectrogram of speech phase does not carry much information about the speech signal, the group delay and IF (which are both obtained from the phase) include similar structures to the speech amplitude. This experimentally proves the fact that speech phase contains useful information about speech signal that can be employed to recover clean speech.

In the following, we present a brief survey of the main directions that have been recently explored about the estimation of speech spectral phase according to [118]. These directions include real-time iterative spectrogram inversion (RTISI), sinusoidal modelling of speech phase, group delay and transient processing, and joint estimation of speech phase and amplitude. The real-time iterative spectrogram inversion (RTISI) is among the first techniques proposed for speech phase estimation. A detailed discussion of this techniques along with a few contributions to each of its steps have been provided in [120]. As Fig. 6 from [120] shows, this method, which is actually based on iterative phase estimation from amplitude, consists of applying STFT synthesis and analysis iteratively with retaining the updated phases and then substituting the updated amplitudes by the given ones. The reasoning behind this approach is to exploit the correlations across neighboring STFT time frames to obtain an estimate of the spectral phase along with the time domain speech signal. Research in this direction has been going on and various contributions have been proposed on modifying the original technique to more efficient methods. As such, in [121], the idea of multiple input spectrogram inversion (MISI) has been proposed, where multiple signals can be reconstructed from their amplitude spectrograms and their mixture signal. Therein, it is shown that the signal spectral phase is a very useful side information which can be employed by imposing that the reconstructed complex spectrograms add up to the mixture complex spectrogram when estimating their phases, resulting in better signal reconstruction quality. Another recent contribution in this direction is to extend the MISI method to modify both the speech amplitude and phase [122]. The latter has led to the informed source separation using iterative reconstruction (ISSIR) method, which is specifically efficient in the context of informed speech source separation, where a quantized version of the oracle amplitude spectrogram is available.

Another main direction in this area is the method of sinusoidal model-based phase estimation presented recently in [123]. Contrary to the aforementioned iterative approaches, this method does not require any estimate of the clean speech amplitude, and instead, the clean spectral phase is estimated by using only an estimate of the fundamental frequency that is blindly obtained from the noisy speech. However, the drawback is that since the sinusoidal modelling is reasonable only for voiced sounds, this approach cannot provide valid estimates of the spectral phase for unvoiced sounds such as fricatives or plosives [118].

Inherent speech information in the phase are not limited to voiced sounds, but are also present for other sounds, like impulses or transients, i.e. sounds of short duration and speech onsets. The speech structures in these sounds appear well in the group delay, as studied in [118], proving that group delay is a useful tool for phase processing [118]. In this sense, group delay has been employed in [124] as a means of performing phase-sensitive noise reduction. Therein, the authors propose to combine a group-delay based phase estimator with a phase-aware amplitude estimator in a closed loop design which checks on the consistency of the estimated amplitude and phase. Other than the group delay,

the IF, corresponding to the temporal derivative of the phase, has also been employed in the literature for the detection of transient sounds. In [125] and references therein, by making use of the fact that the IF changes abruptly when a transient is occurred, algorithms have been proposed for the detection of transient sounds. These algorithms, however, are not only useful in the detection but also in the reduction of transient noises. According to [118], in the presence of transient noise with low SNR, the noisy phase is close to the approximately linear phase of the transient noise. This results in large artifacts if only a amplitude estimator is used along with the noisy phase. To overcome this problem, recently in [126], it has been proposed to use the idea of phase randomization. In this algorithm, the time-frequency units which are dominated by strong transient noise are first detected through the phase-based detection of transient sounds, e.g. in [125]. Next, the phase of theses time-frequency units is replaced by a uniformly distributed random phase, a.k.a., the phase randomization. This helps reducing the effect of the predominant linear phase of the transient noise.

Considering the joint estimation of the spectral amplitude and phase, the first recent work in this direction has been proposed in [127] in the context of Wiener filtering. Contrary to a classical Wiener filter which only aims at the modification of the spectral amplitude, in [127], the relationship among STFT coefficients across time and frequency is considered in order to derive a Wiener filter which modifies both the amplitude and phase of the noisy speech. In this method, under Gaussian assumptions, a joint optimization on phase and amplitude is formed and solved through a conjugate gradient method. Performance assessments show an improved source separation capability as compared to a few previous methods including the classical Wiener filter.

Another recent approach to jointly estimate the speech amplitude and phase has been presented in [128]. Therein, a joint MMSE-based estimator of amplitude and phase is derived directly in the STFT domain, given that an uncertain initial phase estimate is available. This leads to a phase-aware complex estimator of STFT coefficients referred to as the complex estimator with uncertain phase (CUP). The key idea in the derivation of CUP in [128] is to incorporate prior knowledge about speech phase by using a Bayesian framework for phase estimation. In this context, ML and MAP estimators for the spectral phase are derived, assuming complex Gaussian distribution and Chi distribution respectively for noise STFT coefficients and speech amplitudes. Whereas the ML estimator is simply equivalent to the noisy phase of speech, the MAP estimator allows for the incorporation of the prior knowledge of the phase. Furthermore, a joint MMSE-based estimator of the clean speech amplitude and phase is derived, given uncertain prior knowledge of the speech phase. It is shown that while combining a deterministic speech phase estimate with the amplitude may result in annoying artifacts in the enhanced speech (as investigated in [129]), incorporating the uncertainty of the prior phase estimate using the proposed Bayesian estimators reduces these artifacts. Other interesting aspects of this approach include its capability to be used as a statistically optimal phase estimator at the output of an MVDR beamformer and also applicability in a multi-speaker scenario when a multi-pitch tracker is employed to estimate the prior phase information.

Apart from the investigation of different methods of phase estimation, there exists a question of how an available standalone spectral phase estimate can be employed best to improve speech quality. While it may seem that the most obvious way to do this is to combine the separately enhanced amplitude with the estimated phase, according to Wang and Lim [14], combining an independently estimated amplitude and phase does not result in improvements. This can be justified due to the inconsistency between the independently improved amplitude and phase. Instead, a phase-aware amplitude estimator, e.g. [129], should be used in which the amplitude estimator is derived based on an available phase estimate. In [129], it is also demonstrated that the spectral phase can be employed to derive an improved version of speech amplitude estimator being capable of reducing noise outliers that are neglected by the underlying noise PSD estimator. This idea has recently been exploited in [130] in order to propose amplitude estimators that, contrary to the conventional amplitude estimators, treat the spectral phase as a deterministic parameter. Based on an available estimate of the spectral phase, the amplitude estimator is then derived and it is shown that the suggested estimator has the potential to provide even further improvements given more accurate estimates of the spectral phase. Thus, an efficient way to combine a phase estimate with the corresponding enhanced speech amplitude is to use the available phase estimate in the amplitude estimator given by [130] and then reconstruct the enhanced signal using the available phase and obtained amplitude. Another method for combining amplitude and phase estimates is described in [124] where it is proposed to place a phase-aware amplitude estimator in the closed loop of an iterative approach. This approach is enable to enforce consistency within only a few iterations.

As a final conclusion to this section, it should be remarked that speech phase processing is an exciting and

newly emerged field of research, which is capable of extending further the current limits on the performance of amplitude estimators in the STFT domain and make them more robust in challenging acoustic environments.

E. Estimation of Noise PSD Matrix

The noise reduction performance of all beamforming techniques studied in Section II-B1 depends on the accuracy in the estimation of the underlying noise PSD matrix to a large extent. Beamformers such as MVDR and multi-channel Wiener filter use the noise spatial information contained in the noise PSD matrix to be adaptively steered in the direction of interest and reduce the effect of noise impinging on the array from other directions. An inaccurate estimation of the noise PSD matrix can result in unsuccessful cancellation or even amplification of noise directions and also annoying distortion in the enhanced speech. Therefore, for the spatial beamformers to be able to efficiently adapt to the surrounding noise field, accurate knowledge of the noise PSD matrix is necessary [58]. As discussed in Section II-C, there are numerous method for the estimation of the noise PSD in the case of single channel. Yet, until almost recently, the estimation of the noise PSD matrix in a generic non-stationary scenario was not well explored. It is only within the past few years that growing research has been initiated and continuing in this direction. In this section, we discuss in brief the most important approaches to noise PSD matrix estimation with a focus on the most recent methods. We use the same notations used in Section II-B.

By its definition, the noise PSD (or namely correlation) matrix, $\Sigma_{\mathbf{VV}}$, is

$$E\{\mathbf{V}\mathbf{V}^H\} = \begin{bmatrix} E\{|V_1|^2\} & E\{V_1V_2^*\} & \dots & E\{V_1V_N^*\} \\ E\{V_2V_1^*\} & E\{|V_2|^2\} & \dots & E\{V_2V_N^*\} \\ \vdots & \vdots & \ddots & \vdots \\ E\{V_NV_1^*\} & E\{V_NV_2^*\} & \dots & E\{|V_N|^2\} \end{bmatrix} \quad (41)$$

As it is deduced from (41), here, the problem of noise PSD estimation can be extended to the noise PSD matrix estimation with its diagonal elements as the noise auto-PSDs, $E\{|V_i|^2\}$, at individual channels and the non-diagonal elements as the complex noise cross-PSDs, $E\{V_iV_j^*\}$, between each two microphones. The extension of noise PSD estimation techniques to the noise cross-PSD estimation, however, is not often straightforward. This is due to the fact that, unlike noise PSD, the cross-PSD terms are not real-valued and following conventional smoothing, minimum tracking or soft-decision based noise estimators is not reasonably accurate. Despite this, a few efforts have been made rather recently to directly extend/modify the basic noise PSD estimation methods to handle the estimation of the cross-PSD. In this sense, the following smoothing scheme is often used in a soft-decision context to update each of the cross-PSD terms

$$\begin{aligned} \hat{\sigma}_{v_iv_j}(k,l) = (1-\alpha_s(k,l))\hat{\sigma}_{v_iv_j}(k,l-1) \\ + \alpha_s(k,l)X_i(k,l)X_j^*(k,l) \end{aligned} \quad (42)$$

with $\hat{\sigma}_{v_iv_j}(k,l)$ denoting the estimate of the cross-PSD term $E\{V_iV_j^*\}$, $\alpha_s(k,l)$ as the smoothing parameter, and $X_i(k,l)$ and $X_j(k,l)$ denoting the noisy observations in the ith and jth channels respectively. In [131], by using a joint VAD for both of the two channels, presence of speech is detected at each time-frequency unit and then the basic MS method along with (42) are used to update the noise cross-PSDs in time-frequencies where speech does not exist. Yet, since the noise cross-PSDs are updated only at speech pauses, this approach cannot trace fast changes in a non-stationary noisy field. In general, denoting by $\mathcal{H}_0(k,l)$ and $\mathcal{H}_1(k,l)$ respectively the states where absence and presence of speech is detected by a VAD, the VAD-based cross-PSD estimators act as the following scheme

$$\hat{\sigma}_{v_iv_j}(k,l) = \begin{cases} \text{update } \hat{\sigma}_{v_iv_j}(k,l), & \text{under } \mathcal{H}_0(k,l) \\ \hat{\sigma}_{v_iv_j}(k,l-1), & \text{under } \mathcal{H}_1(k,l) \end{cases} \quad (43)$$

where the updating of $\hat{\sigma}_{v_iv_j}(k,l)$ is performed by an extension of a noise PSD estimation method. Later in [132], a more advanced extension of the MS approach for noise PSD estimation was presented and then extended to cross-PSD estimation. The entire algorithm is soft-decision based and no VADs are used. Performance comparisons of the suggested approach to two other noise cross-PSD estimators based on VAD and MS revealed improvements in noise reduction.

Another group of methods for noise PSD matrix estimation assume certain structures for the type of noise field. The most important methods in this group are those considering a diffuse noise field which is based on the fact that for the case of a spherically or cylindrically isotropic diffuse noise field, the coherence function, $\Gamma_{v_iv_j}(k,l)$, is ideally known and depends only on the frequency and the distances between microphones, as the following [133]

$$\Gamma_{v_iv_j}(k,l) \triangleq \frac{\sigma_{v_iv_j}(k,l)}{\sqrt{\sigma_{v_i}^2(k,l)\sigma_{v_j}^2(k,l)}} = \sin c\left(\frac{\omega f_s d_{i,j}}{C}\right) \quad (44)$$

with ω the angular frequency defined as $2\pi k/K$, f_s the sampling frequency, $d_{i,j}$ the distance between microphones i and j, and C as the sound velocity. However, imposing a diffuse noise field assumption is not always realistic and these coherence-based methods cannot be used for the general scenario of spatially (across microphones) highly correlated noise. To improve the

performance of such methods, a two-microphone approach for the estimation of noise cross-PSD based on the assumption of a diffuse noise field has been proposed in [134]. In this approach the speech phase information has been employed to estimate the noise cross-PSD, which in turn, is used to calculate a coherence-based gain function for noise reduction. In the same direction, the same authors present another two-microphone method in [135] where the noise cross-PSD and a noise reduction filter gain are iteratively estimated. The filter gain is used to mitigate the speech components in the estimated noise cross-PSD in order to avoid leakage of speech into the noise estimates. Later in [133], instead of directly estimating the noise cross-PSDs and using them in a beamformer, a dual-microphone speech enhancement technique is suggested based on the coherence function. Without requiring to estimate noise statistics directly, this technique utilizes the coherence function between the speech and noise sources as a criterion for noise reduction and formulated a filter whose coefficients are dependent on the estimated coherence function. The suggested method was evaluated particularly in a dual-microphone application with highly spatially correlation noise and yielded a substantial improvement with respect to conventional beamforming. More recently in [136], another dual-microphone approach has been proposed, which is particularly useful in a mobile phone in hands-free position. In this work, instead of using predefined models for the coherence function, a single microphone noise PSD estimation algorithm based on the SPP and a dual microphone technique exploiting the coherence properties of the speech source and the background noise are combined. Therein, a new technique is presented to estimate the coherence function which is not practically known.

The idea of the SPP has primarily been used in the context of noise PSD estimation particularly by the famous IMCRA method of Cohen [80]. Souden et al. extend this idea to the multi-channel case in [137] based on Gaussian models of speech and noise. The principle contribution in [137] is to extend the definition of the SPP from $\mathcal{P}\{\mathcal{H}_1 \mid X(k,l)\}$ to $\mathcal{P}\{\mathcal{H}_1 \mid X(k,l)\}$ for the multi-channel case in a Bayesian framework. The authors use a fixed *a priori* SPP, $\mathcal{P}\{\mathcal{H}_0\}$, and derive a closed-form expression for the multi-channel SPP under the assumption that the speech and noise components are complex multivariate Gaussian and that the real and imaginary parts of all signals are uncorrelated and identically distributed. It is shown by theoretical and numerical evaluations that the suggested multi-channel SPP increases the detection accuracy of speech as compared to the classical single-channel SPP. Later in [138], the authors employ the proposed multi-channel SPP along with an alternative formulation of the IMCRA method to propose a recursive algorithm for the noise PSD matrix estimation. Therein, the multi-channel SPP is employed for the accurate detection of speech components and then an iterative modification of the IMCRA approach is performed for tracking the noise PSD matrix. The ith iteration of this algorithm can be summarized as the following [138]

$$\begin{aligned}\Sigma_{\mathbf{VV}}^{(i-1)}(k,l) \;&\mapsto\; \mathcal{P}\{\mathcal{H}_1 \mid \mathbf{X}(k,l)\}^{(i)}\\ \alpha_v^{(i)}(k,l) &= \alpha_{v0} + (1-\alpha_{v0})\mathcal{P}\{\mathcal{H}_1 \mid \mathbf{X}(k,l)\}^{(i)}\\ \Sigma_{\mathbf{VV}}^{(i)}(k,l) &= \alpha_v^{(i)}(k,l)\Sigma_{\mathbf{VV}}^{(i)}(k,l-1) + (1-\alpha_v^{(i)}(k,l))\\ &\qquad \mathbf{X}(k,l)\mathbf{X}^H(k,l)\end{aligned} \tag{45}$$

As it is observed, the estimated multi-channel SPP, i.e. $\mathcal{P}\{\mathcal{H}_1 \mid X(k,l)\}^{(i)}$, which is used to obtain the smoothing parameter $\alpha_v^{(i)}(k,l)$ in the second line, is itself a function of the noise PSD matrix $\Sigma_{\mathbf{VV}}^{(i-1)}(k,l)$, as indicated in the first line. Thus, a recursion exists between the SPP and the noise PSD matrix, which has to be conducted until convergence is reached. The performance of the proposed algorithm is assessed by using it in different beamformers for noise reduction in various conditions including stationary or non-stationary noise and in anechoic or reverberant acoustic rooms. It is demonstrated that good performance in terms of speech detection, noise tracking and noise reduction is obtained. As well, the proposed multi-channel SPP in [137] has been more recently employed in [139] in order to derive an improved multi-channel Wiener filter. The latter, as compared to using the noise PSD matrix estimation in [138] in beamforming, helps significantly reduce the background noise while suffers from only a little speech distortion.

Hendriks and Gerkmann investigate the noise PSD matrix estimation problem from another aspect in [140]. Therein, they present a general approach which can be applied to a non-stationary noise scenario without adopting a VAD or a coherence function for the noisy observations or considering any assumptions about the distribution of noise or speech. Rather, they exploit the fact that if the steering vector, i.e. $\mathbf{A}$ in (16), of the speech source is known, a noise reference can be calculated, that is independent of any speech components. Specifically, the noise reference between the two microphones i and j is given as [140]

$$P_{i,j} = X_i - \frac{A_i}{A_j}X_j = V_i - \frac{A_i}{A_j}V_j \tag{46}$$

with A_i as the ith element of the steering vector $\mathbf{A}$. This noise reference can be exploited to estimate the cross-PSD term, based on the conventional assumption that speech and noise are uncorrelated and that there is

ideally no speech component in $P_{i,j}$ in (46). In this regard, the following expression can be derived [140]

$$\hat{\sigma}_{v_i v_j} = E\{P_{i,j} X_j^*\} + \frac{A_i}{A_j}\sigma_{v_j}^2 \tag{47}$$

Therefore, the estimation of the cross-PSD term, $\hat{\sigma}_{v_i v_j}$, reduced to the estimation of $E\{P_{i,j}X_j^*\}$ and $\sigma_{v_j}^2$, which are, respectively, estimated by a simple recursive smoothing and a single-channel noise PSD estimator from the literature. However, both of these estimation procedures are not error-free and they reduce the accuracy of the proposed method. The authors suggest to use the Hermitian symmetry property of the PSD matrix and calculate the ultimate estimate of $\sigma_{v_i v_j}$ as the average of the two terms $\hat{\sigma}_{v_i v_j}$ and $(\hat{\sigma}_{v_j v_i})^*$ obtained from (47). Moreover, it is shown that if the proposed noise PSD matrix estimator is employed in an MVDR beamformer under far-field and free-field conditions, the ultimate form of the MVDR weights become independent of the underlying noise auto-PSD estimates and therefore, the estimation error decreases even further. Performance of the suggested noise PSD matrix estimation approach is evaluated by employing it in an MVDR beamformer in noisy and reverberant environments and measuring different quality objectives and the superiority of the suggested method is proved with respect to a few other method such as VAD-based noise estimation and the GSC method. However, it should be noted that perfect knowledge of the steering vector A is assumed in this approach, and this assumption, particularly in reverberant environments, is not quite realistic. Since in such environments, this assumption is equivalent to knowing the transfer function of the acoustic room which is not often available. The main idea in [140], which is based on blocking the signal components in noisy observations prior to the calculation of the noise PSD matrix, has been also exploited in [141]. Therein, a blocking matrix, similar to that in the GSC method, is used to mitigate the speech components and the resulting output is used in an ML framework to estimate the noise PSD matrix. Evaluation of this method shows satisfying performance for high SNR values where the speech component is dominant.

As two more recently proposed approaches to noise PSD matrix estimation in a generic non-stationary noise field with no limiting assumptions, the works in [142] and [143] can be mentioned. The work in [142], which is based on the popular IMCRA method consists of two main contributions. The first contribution is an improvement to the single-channel IMCRA method, where a special noise level detector is employed in order to enhance the noise tracking capability of the original IMCRA. The second contribution concerns the estimation of the noise cross-PSD by means of a smoothed cross-periodogram. The latter is obtained by using estimated noise-only components derived as residuals after applying speech enhancement on the noisy observations at each channel. Evaluation of the suggested approach shows its advantage when used by an MVDR beamformer in noisy and also reverberant environments. As well, the robust noise PSD matrix estimation approach presented in [143] assumes a non-stationary noise field without prior knowledge about the noise or speech. In this approach, a smoothing scheme for the noise PSD estimation is proposed, which takes advantage of the close subsequent speech frames in addition to the current and past frames. The smoothing parameter in the proposed smoothing scheme is calculated in as a function of an overall SNR measure in all channels. Since the latter is obtained based on an available estimate of the noise PSD matrix, similar to [138], the smoothing parameter becomes a function of the noise PSD matrix and thus a recursive algorithm is formulated. As a second stage, an extension of the MS approach is applied over the primary estimate of the noise PSD matrix obtained from the recursion, in order to increase the noise estimation accuracy. The proposed recursive method converges within only two iterations, and thanks to its increased accuracy in the second stage, it outperforms two other recent noise PSD matrix estimation methods in the literature.

IV. Summary and Conclusions

This work presented an overview on different aspects of noise reduction methods in the STFT domain. In general, the straightforward implementation and low computational costs have made these methods appealing for practical and real-time applications. In Section II, a brief review of the conventional methods in this field was presented. In the case of single-channel approaches, we studied in brief spectral subtractive methods, Wiener filtering based methods, estimators of speech STSA including Bayesian (MMSE) and MAP approaches and estimators of the complex speech STFT coefficients. In the case of multi-channel speech enhancement, we briefly reviewed the most important conventional beamforming and post-filtering techniques. The former includes DAS beamformer, multi-channel Wiener filter and its distortionless version, maximum SNR spatial filter and the MVDR beamformer; and the latter includes Zelinski's post-filter, Wiener-based and coherence-based post-filters, and a special post-filtering method for the GSC beamformer. This section is followed by a brief review on noise estimation methods and performance measures for speech enhancement.

Section III discussed the most recent contributions in STFT-based noise reduction algorithms. In subsection A, the use of super-Gaussian distributions and the

estimation of their parameters to model the speech prior in STSA estimators were investigated. It is notable that all these distributions are special cases of the GGD, and therefore, the underlying STSA estimators differ mostly in the parameter selection of the speech prior. It can be concluded that whereas the investigation of more sophisticated Bayesian cost functions (than those presented in Table 2) has not been considered much for the past few years, there has been growing interest in the literature in the employment of more perceptually meaningful prior distributions with adaptation of their parameters to the speech STSA. In this regard, since there has not been an optimal scheme for the parameter selection nor a unified criterion for the adaptation of the speech prior parameters, further research is required to enhance or optimize the performance of STSA estimators with parametric speech priors.

Subsection II-C covered the most famous methods of noise PSD estimation, which include VAD-based methods, MS tracking, IMCRA, and the more recent statistical model-based approaches such as MMSE, MAP and ML methods. Contrary to the hard-decision (i.e. VAD-based) methods, which tend to estimate the noise PSD only in the absence of speech components, the soft-decision methods are able to potentially update noise PSD estimates in all speech frames. It is well-known that the soft-decision group of methods outperforms the hard-decision group. Despite this, still, approaches involving a combination of the two group of methods receives attention in the literature. Also, various attempts have been made to improve the empirical schemes and experimentally chosen parameters of the MS and IMCRA approaches. In doing so, decreasing the estimation error variance and a faster tracking of abrupt noise changes have been targeted the most. Even though the performance of noise estimation methods depends to a high degree on the noise conditions, it appears that the recently proposed statistical model-based methods, especially the MMSE-based approach, are able to offer good performance with more robustness to adverse non-stationary noise conditions.

Another appealing direction recently suggested is the spectral subtraction in the modulation domain which was reviewed in subsection III-C. It is concluded that there exists possibility to improve the performance of the STFT-based noise reduction methods by implementing them in the modulation domain at the cost of only an additional Fourier and inverse Fourier transforms. Also, fusion of this domain with the conventional STFT domain has been shown to be helpful in establishing a trade-off between the drawbacks of each domain. So far, various modifications and improvements have been applied to this approach, most of which have been borrowed from the conventional STFT domain.

Speech phase processing in the STFT domain is another topic which has, in the past few years, drawn lots of interest in the literature. This topic was studied in subsection D. Major phase estimation approaches proposed to date include real-time iterative spectrogram inversion, sinusoidal phase modeling, group delay and transient processing, and joint amplitude-phase estimation. The latter has been derived in a Bayesian framework and is more similar to the conventional amplitude estimators in the implementation. Based on extensive recent investigations, it is proved that using phase estimates instead of the noisy phase can provide further quality and intelligibility improvements, given that the estimated phase is properly combined with the estimated amplitude, e.g. used in a phase-aware amplitude estimator. However, as compared to speech amplitude estimation, phase estimation seems to be a more complicated problem with still many aspects to explore. Yet, with the increase in technology and the processing power, phase processing is an exciting area of research that is likely to lead to further push in the current limits on speech enhancement.

Finally in subsection E, estimation of the noise PSD matrix was investigated. Whereas numerous methods have been introduced for the noise PSD estimation, not as many have been suggested for the estimation of the noise PSD matrix so far. Compared to the noise PSD estimation, the problem here is more challenging as the cross-PSD is generally complex-valued and noise can also be spatially non-stationary. Yet, recently, this topic has captured more attention by trying to improve and then extend many single-channel noise estimation methods to their multi-channel counterpart. A few extensions, in this regard, have been presented to estimate noise cross-PSD terms and then combine them with the corresponding auto-PSD terms to built an estimation of the noise PSD matrix consistently. Other than these extensions, a few important methods include coherence-based approaches, soft-decision methods based on multi-channel SPPs, the method of Hendriks and Gerkmann, and a few more recent combinational approaches. Still, with the large performance gap between the ideal noise PSD matrix estimation, i.e. by using noise-only samples, and the suggested methods, there is considerable room for further research in this newly explored topic.

Mahdi Parchami received the B.Sc. degree in electrical engineering from Shahed University, Tehran, Iran, in 2006, and the M.Sc. degree in electrical engineering (communications) from Amirkabir University of Technology (Tehran Polytechnic), Iran, in 2009. Since 2011, he has been with the department of electrical and computer engineering at Concordia University, Montreal, QC, Canada, where

he works with the signal processing group as a research assistant toward the Ph.D. degree. His research area includes adaptive signal processing, detection and estimation, MIMO acoustic signal processing, and speech enhancement with focus on noise reduction and reverberation suppression.

Wei-Ping Zhu (SM'97) received the B.E. and M.E. degrees from Nanjing University of Posts and Telecommunications, and the Ph.D. degree from Southeast University, Nanjing, China, in 1982, 1985, and 1991, respectively, all in electrical engineering. He was a Postdoctoral Fellow from 1991 to 1992 and a Research Associate from 1996 to 1998 with the Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada. During 1993–1996, he was an Associate Professor with the Department of Information Engineering, Nanjing University of Posts and Telecommunications. From 1998 to 2001, he worked with hi-tech companies in Ottawa, Canada, including Nortel Networks and SR Telecom Inc. Since July 2001, he has been with Concordia's Electrical and Computer Engineering Department as a full-time faculty member, where he is presently a Full Professor. His research interests include digital signal processing fundamentals, speech and statistical signal processing, and signal processing for wireless communication with a particular focus on MIMO systems and cooperative communication.

Dr. Zhu served as an Associate Editor for the IEEE Transactions on Circuits and Systems Part I: Fundamental Theory and Applications during 2001-2003, an Associate Editor for Circuits, Systems and Signal Processing during 2006-2009, and an Associate Editor for the IEEE Transactions on Circuits and Systems Part II: Transactions Briefs during 2011-2015. He was also a Guest Editor for the IEEE Journal on Selected Areas in Communications for the special issues of: Broadband Wireless Communications for High Speed Vehicles, and Virtual MIMO during 2011-2013. Currently, he is an Associate Editor of Journal of The Franklin Institute. Dr. Zhu was the Chair-Elect of Digital Signal Processing Technical Committee (DSPTC) of the IEEE Circuits and System Society during June 2012-May 2014, and the Chair of the DSPTC during June 2014-May 2016.

Benoit Champagne received the B.Ing. degree in Engineering Physics from the École Polytechnique de Montréal in 1983, the M.Sc. degree in Physics from the Université de Montréal in 1985, and the Ph.D. degree in Electrical Engineering from the University of Toronto in 1990. From 1990 to 1999, he was an Assistant and then Associate Professor at INRS-Telecommunications, Université du Quebec, Montréal. In 1999, he joined McGill University, Montreal, where he is now a Full Professor in the Department of Electrical and Computer Engineering; he also served as Associate Chairman of Graduate Studies in the Department from 2004 to 2007. His research focuses on the study of advanced algorithms for the processing of communication signals by digital means. His interests span many areas of statistical signal processing, including detection and estimation, sensor array processing, adaptive filtering, and applications thereof to broadband communications and audio processing, where he has co-authored nearly 250 referred publications. His research has been funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada, the "Fonds de Recherche sur la Nature et les Technologies" from the Govt. of Quebec, as well as some major industrial sponsors, including Nortel Networks, Bell Canada, InterDigital and Microsemi.

He has been an Associate Editor for the *EURASIP J. on Applied Signal Processing* from 2005 to 2007, the *IEEE Signal Processing Letters* from 2006 to 2008, and the *IEEE Trans. on Signal Processing* from 2010 to 2012, as well as a Guest Editor for two special issues of the *EURASIP J. on Applied Signal Processing* published in 2007 and 2014, respectively. He has also served on the Technical Committees of several international conferences in the fields of communications and signal processing. In particular, he was Registration Chair, for IEEE ICASSP 2004, Co-Chair, Antenna and Propagation Track, for IEEE VTC–Fall 2004, Co-Chair, Wide Area Cellular Communications Track, for IEEE PIMRC 2011, Co-Chair, Workshop on D2D Communications, for IEEE ICC 2015 and Publicity Chair, for IEEE VTC–Fall 2016. His is currently a Senior Member of IEEE.

Eric Plourde received both the B.Ing. degree in electrical engineering and the M.Sc.A. degree in biomedical engineering from the Ecole Polytechnique de Montréal, QC, Canada in 2003. He completed a Ph.D. degree in electrical engineering from McGill University, Montreal, QC, Canada in 2009. From 2009 to 2011, he was a Postdoctoral Fellow in the Neuroscience Statistics Research Laboratory with joint appointments at the Massachusetts General Hospital, Harvard Medical School and the Massachusetts Institute of Technology. He joined the Université de Sherbrooke in 2011, where he is now an Associate Professor within the Department of Electrical and Computer Engineering. His research interests include neural signal processing as well as speech processing with emphasis on speech enhancement and perceptually/biologically inspired processing.

References

[1] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*. Berlin, Heidelberg: Springer, 2005.

[2] I. Cohen, J. Benesty, and S. Gannot, Eds., *Speech Processing in Modern Communication: Challenges and Perspectives*. Berlin, Heidelberg: Springer, 2012.

[3] J. Benesty and Y. Huang, Eds., *Springer Handbook of Speech Processing*. Berlin, Heidelberg: Springer, 2008.

[4] S. Vaseghi, Ed., *Advanced Digital Signal Processing and Noise Reduction*. Chichester, UK: Wiley, 2008.

[5] J. Benesty and Y. Huang, Eds., *Adaptive Signal Processing: Applications to Real-World Problems*. Berlin, Heidelberg: Springer, 2003.

[6] J. Benesty, Ed., *Advances in Network and Acoustic Echo Cancellation*. Berlin, Heidelberg: Springer, 2001.

[7] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC Press, 2007.

[8] S. I. Yann, "*Transform based speech enhancement techniques*," Ph.D. dissertation, School of Elec. & Electron. Eng., Nanyang Technol. Univ., Singapore, 2003.

[9] J. Benesty, J. Chen, and E. Habets, Eds., *Speech Enhancement in the STFT Domain*. Berlin, Heidelberg: Springer, 2011.

[10] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoustics Speech Signal Process.,* vol. 28, no. 2, pp. 137–145, Apr. 1980.

[11] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoustics Speech Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[12] J. Porter and S. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1984, vol. 9, pp. 53–56.

[13] R. Martin, "Speech enhancement based on minimum mean-square error estimation and Supergaussian priors," *IEEE Trans. Speech Audio Process.,* vol. 13, no. 5, pp. 845–856, Sept. 2005.

[14] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoustics Speech Signal Process.*, vol. 30, no. 4, pp. 679–681, Aug. 1982.

[15] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1983, vol. 8, pp. 804–807.

[16] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics Speech Signal Process.,* vol. 32, no. 6, pp. 1109–1121, 1984.

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

[18] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoustics Speech Signal Process.,* vol. 28, no. 2, pp. 137–145, 1980.

[19] S. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Upper Saddle River, NJ: Prentice Hall, 1993. [Database] [Mismatch]

[20] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics Speech Signal Process.,* vol. 33, no. 2, pp. 443–445, 1985.

[21] P. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.,* vol. 13, no. 5, pp. 857–869, 2005.

[22] C. You, S. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 475–486, July 2005.

[23] D. Brillinger, *Time Series: Data Analysis and Theory*. Philadelphia, PA: SIAM, 2001.

[24] A. Oppenheim and J. Lim, "The importance of phase in signals," *Proc. IEEE,* vol. 69, no. 5, pp. 529–541, May 1981.

[25] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.,* vol. 53, no. 4, pp. 465–494, 2011.

[26] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Process. Lett.,* vol. 20, no. 2, pp. 129–132, Feb. 2013.

[27] P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," *IEEE Signal Process. Lett.,* vol. 20, no. 12, pp. 1235–1239, Dec. 2013.

[28] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio Speech, Lang. Process.,* vol. 22, no. 12, pp. 1931–1940, Dec. 2014.

[29] K. Paliwal and D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Commun.,* vol. 45, no. 2, pp. 153–170, 2005.

[30] P. Vary, "Noise suppression by spectral magnitude estimation: mechanism and theoretical limits," *Signal Process.*, vol. 8, no. 4, pp. 387–400, 1985.

[31] H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Speech Audio Process.,* vol. 9, no. 8, pp. 799–807, Nov. 2001.

[32] Y. Hu, M. Bhatnagar, and P. Loizou, "A cross-correlation technique for enhancing speech corrupted with correlated noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, 2001, vol. 8, pp. 673–676.

[33] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Process.,* vol. 7, no. 2, pp. 126–137, Mar. 1999.

[34] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*. Berlin, Heidelberg: Springer, 2007.

[35] B. Sim, Y. Tong, J. Chang, and C. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 328–337, July 1998.

[36] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 334–341, July 2003.

[37] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Commun.,* vol. 50, no. 6, pp. 453–466, 2008.

[38] R. Udrea and S. Ciochina, "Speech enhancement using spectral over-subtraction and residual noise reduction," in *Proc. Int. Symp. Signals, Circuits and Systems*, 2003, vol. 1, pp. 165–168.

[39] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars," *Speech Commun.,* vol. 11, no. 2, pp. 215–228, 1992.

[40] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2002, vol. 4, pp. IV–4164–IV–4164.

[41] P. Sovka, P. Pollack, and J. Kybic, "Extended spectral subtraction," in *Proc. European Signal Processing Conf.*, 1996, pp. 963–966.

[42] Y. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *IEEE Trans. Signal Process.,* vol. 39, no. 9, pp. 1943–1954, Sept. 1991.

[43] S. Haykin, *Adaptive Filter Theory*. Upper Saddle River, NJ: Pearson Education, 2008.

[44] S. Marple, *Digital Spectral Analysis with Applications*. Upper Saddle River, NJ: Prentice-Hall, Inc., 1986.

[45] M. Hasan, S. Salahuddin, and M. Khan, "A modified a priori SNR for speech enhancement using spectral subtraction rules," *IEEE Signal Process. Lett.,* vol. 11, no. 4, pp. 450–453, Apr. 2004.

[46] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Trans. Speech Audio Process.,* vol. 13, no. 5, pp. 870–881, Sept. 2005.

[47] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.,* vol. 14, no. 6, pp. 2098–2108, Nov. 2006.

[48] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.,* vol. 14, no. 1, pp. 163–176, Jan. 2006.

[49] Y. Hu and P. Loizou, "A perceptually motivated approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 457–465, Sept. 2003.

[50]Y. Hu and P. C. Loizou, "Incorporating a psychoacoustical model in frequency domain speech enhancement," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 270–273, Feb. 2004.

[51] E. Plourde and B. Champagne, "Auditory-based spectral amplitude estimators for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.,* vol. 16, no. 8, pp. 1614–1623, Nov. 2008.

[52] E. Plourde and B. Champagne, "Generalized Bayesian estimators of the spectral amplitude for speech enhancement," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 485–488, June 2009.
[53] P. Wolfe and S. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP J. Appl. Signal Process*., vol. 2003, no. 10, pp. 1043–1051, 2003.
[54] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process*., vol. 2005, no. 7, pp. 1110–1126, 2005.
[55] J. Erkelens, R. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio Speech Lang. Process.,* vol. 15, no. 6, pp. 1741–1752, Aug. 2007.
[56] T. Tuncer and B. Friedlander, *Classical and Modern Direction-of-Arrival Estimation*. Burlington, MA: Elsevier Science, 2009.
[57] Z. Chen, G. Gokeda, and Y. Yu, *Introduction to Direction-of-Arrival Estimation*. Norwoor, MA: Artech House, 2010.
[58] J. Benesty, J. Chen, and Y. Huang, Eds., *Microphone Array Signal Processing*. Berlin, Heidelberg: Springer, 2010.
[59] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio Speech Lang. Process*., vol. 18, no. 2, pp. 260–276, 2010.
[60] E. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Trans. Audio Speech Lang. Process.,* vol. 18, no. 1, pp. 158–170, Jan. 2010.
[61] C. Marro, Y. Mahieux, and K. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 240–259, May 1998.
[62] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 1988, vol. 5, pp. 2578–2581.
[63] K. Simmer and A. Wasiljeff, "Adaptive microphone arrays for noise suppression in the frequency domain," in *Proc. Second Cost 229 Workshop on Adaptive Algorithms in Communications*, Bordeaux, France, 1992, pp. 185–194.
[64] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.,* vol. 11, no. 6, pp. 709–716, Nov. 2003.
[65] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech Audio Process.,* vol. 12, no. 6, pp. 561–571, Nov. 2004.
[66] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.,* vol. 9, no. 4, pp. 113–116, Apr. 2002.
[67] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Speech Audio Process.,* vol. 10, no. 6, pp. 341–351, Sept. 2002.
[68] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-the-Art*. San Rafael, CA: Morgan & Claypool, 2013.
[69] S. V. Gerven and F. Xie, "A comparative study of speech detection methods," in *Proc. 5th European. Conf. Speech Communication and Technology*, Rhodes, Greece, 1997, pp. 1–4.
[70] Ó. Varela, R. San-Segundo, and L. A. Hernández, "Combining pulse-based features for rejecting far-field speech in a HMM-based voice activity detector," *Comp. Elec. Eng.,* vol. 37, no. 4, 2011.
[71] T. Fukuda, O. Ichikawa, and M. Nishimura, "Long-term spectro-temporal and static harmonic features for voice activity detection," *IEEE J. Selected Topics Signal Process.,* vol. 4, no. 5, pp. 834–844, Oct. 2010.
[72] D. Vlaj, Z. Kačič, and M. Kos, "Voice activity detection algorithm using nonlinear spectral weights, hangover and hangbefore criteria," *Comp. Elec. Eng.,* vol. 38, no. 6, pp. 1820–1836, 2012.
[73] M. Mak and H. Yu, "A study of voice activity detection techniques for NIST speaker recognition evaluations," *Comp. Speech Lang.,* vol. 28, no. 1, pp. 295–313, 2014.
[74] P. Teng and Y. Jia, "Voice activity detection via noise reducing using non-negative sparse coding," *IEEE Signal Process. Lett.*, vol. 20, no. 5, pp. 475–478, May 2013.
[75] S. Mousazadeh and I. Cohen, "Voice activity detection in presence of transient noise using spectral clustering," *IEEE Trans. Audio Speech Lang. Process.,* vol. 21, no. 6, pp. 1261–1271, June 2013.
[76] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.,* vol. 9, no. 5, pp. 504–512, July 2001.
[77] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Process.*, vol. 86, no. 6, pp. 1215–1229, 2006.
[78] N. Derakhshan, A. Akbari, and A. Ayatollahi, "Noise power spectrum estimation using constrained variance spectral smoothing and minima tracking," *Speech Commun.,* vol. 51, no. 11, pp. 1098–1113, 2009.
[79] J.-H. Chang, "Noisy speech enhancement based on improved minimum statistics incorporating acoustic environment-awareness," *Digital Signal Process.*, vol. 23, no. 4, pp. 1233–1238, 2013.
[80] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.,* vol. 11, no. 5, pp. 466–475, Sept. 2003.
[81] N. Fan, J. Rosca, and R. Balan, "Speech noise estimation using enhanced minima controlled recursive averaging," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2007, vol. 4, pp. IV–581–IV–584.
[82] B. Borgstrom and A. Alwan, "Improved speech presence probabilities using HMM-based inference, with applications to speech enhancement and ASR," *IEEE J Selected Topics Signal Process.,* vol. 4, no. 5, pp. 808–815, Oct. 2010.
[83] W. Yuan, J. Lin, W. An, Y. Wang, and N. Chen, "Noise estimation based on time–frequency correlation for speech enhancement," *Appl. Acoust.,* vol. 74, no. 5, pp. 770–781, 2013.
[84] H. Momeni, E. Habets, and H. Abutalebi, "Single-channel speech presence probability estimation using inter-frame and inter-band correlations," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2014, pp. 2903–2907.
[85] B. Fodor and T. Gerkmann, "A posteriori speech presence probability estimation based on averaged observations and a super-Gaussian speech model," in *Proc. 14th Int. Workshop on Acoustic Signal Enhancement*, 2014, pp. 11–15.
[86] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.,* vol. 16, no. 1, pp. 229–238, Jan. 2008.
[87] P. E. Papamichalis, *Practical Approaches to Speech Coding*. New York: Prentice-Hall, 1987.
[88] L. Weisi, D. Tao, J. Kacprzyk, Z. Li, E. Izquierdo, and H. Wang, Eds., *Multimedia Analysis, Processing and Communications*. Berlin Heidelberg: Springer, 2011.
[89] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2001, vol. 2, pp. 749–752.
[90] B. Grundlehner, J. Lecocq, R. Balan, and J. Rosca, "Performance assessment method for speech enhancement systems," in *Proc. 1st Annu. IEEE BENELUX/DSP Valley Signal Process. Symp.*, 2005, pp. 1–4.
[91] B. Fodor, "*Contributions to statistical modeling for minimum mean square error estimation in speech enhancement*," Ph.D. dissertation, Technische Universität, Braunschweig, 2015.
[92] J. W. Shin, J.-H. Chang, and N. S. Kim, "Statistical modeling of speech signals based on generalized Gamma distribution," *IEEE Signal Process. Lett.*, vol. 12, no. 3, pp. 258–261, Mar. 2005.
[93] R. Prasad, H. Saruwatari, and K. Shikano, "Probability distribution of time-series of speech spectral components," *IEICE Trans. Fundamentals Electron. Commun, Comp. Sci.*, vol. E87-A, no. 3, pp. 584–597, Mar. 2004.
[94] B. Borgstrom and A. Alwan, "A unified framework for designing optimal STSA estimators assuming maximum likelihood phase equivalence of speech and noise," *IEEE Trans. Audio, Speech Lang. Process.,* vol. 19, no. 8, pp. 2579–2590, Nov. 2011.
[95] B. J. Borgstrom and A. Alwan, "Log-spectral amplitude estimation with generalized Gamma distributions for speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Prague, 2011, pp. 4756–4759.
[96] I. Andrianakis and P. White, "Speech spectral amplitude estimators using optimally shaped Gamma and Chi priors," *Speech Commun.,* vol. 51, no. 1, pp. 1–14, 2009.
[97] Y.-C. Su, Y. Tsao, J.-E. Wu, and F.-R. Jean, "Speech enhancement using generalized maximum a posteriori spectral amplitude estimator," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2013, pp. 7467–7471.
[98] M. B. Trawicki and M. T. Johnson, "Speech enhancement using Bayesian estimators of the perceptually-motivated short-time spectral

amplitude (STSA) with Chi speech priors," *Speech Commun.,* vol. 57, pp. 101–113, 2014.
[99] H. R. Abutalebi and M. Rashidinejad, "Speech enhancement based on β-order MMSE estimation of short time spectral amplitude and Laplacian speech modeling," *Speech Commun.,* vol. 67, pp. 92–101, Mar. 2015.
[100] B. Chen and P. Loizou, "A Laplacian-based MMSE estimator for speech enhancement," *Speech Commun.,* vol. 49, no. 2, pp. 134–143, 2007.
[101] M. Parchami, W.-P. Zhu, B. Champagne, and E. Plourde, "Bayesian STSA estimation using masking properties and generalized Gamma prior for speech enhancement," *EURASIP J. Adv. Signal Process.,* vol. 2015, no. 1, p. 87, 2015.
[102] O. Gomes, C. Combes, and A. Dussauchoy, "Parameter estimation of the generalized Gamma distribution," *Math. Comp. Simul.,* vol. 79, no. 4, pp. 955–963, 2008.
[103] A. Chinaev, A. Krueger, D. H. T. Vu, and R. Haeb-Umbach, "Improved noise power spectral density tracking by a MAP-based postprocessor," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2012, pp. 4041–4044.
[104] A. Chinaev, R. Haeb-Umbach, J. Taghia, and R. Martin, "Improved single-channel nonstationary noise tracking by an optimized MAP-based postprocessor," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2013, pp. 7477–7481.
[105] R. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, 2010, pp. 4266–4269.
[106] T. Gerkmann and R. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
[107] T. Gerkmann and R. C. Hendriks, "Improved MMSE-based noise PSD tracking using temporal cepstrum smoothing," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2012, pp. 105–108.
[108] M. Souden, M. Delcroix, K. Kinoshita, T. Yoshioka, and T. Nakatani, "Noise power spectral density tracking: A maximum likelihood perspective," *IEEE Signal Process. Lett.,* vol. 19, no. 8, pp. 495–498, Aug. 2012.
[109] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2011, pp. 4640–4643.
[110] R. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio, Speech Lang. Process.,* vol. 16, no. 3, pp. 541–553, Mar. 2008.
[111] K. Paliwal, B. Schwerin, and K. Wójcicki, "Role of modulation magnitude and phase spectrum towards speech intelligibility," *Speech Commun.,* vol. 53, no. 3, pp. 327–339, 2011.
[112] L. Atlas, Q. Li, and J. Thompson, "Homomorphic modulation spectra," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2004, vol. 2, pp. ii–761–4.
[113] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Commun.,* vol. 52, no. 5, pp. 450–475, 2010.
[114] K. Paliwal, B. Schwerin, and K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Commun.,* vol. 54, no. 2, pp. 282–305, 2012.
[115] K. K. Wójcicki and P. C. Loizou, "Channel selection in the modulation domain for improved speech intelligibility in noise," *J. Acoust. Soc. Am.,* vol. 131, no. 4, pp. 2904–2913, 2012.
[116] S. So and K. K. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement," *Speech Commun.*, vol. 53, no. 6, pp. 818–829, 2011.
[117] Y. Zhang and Y. Zhao, "Real and imaginary modulation spectral subtraction for speech enhancement," *Speech Commun.,* vol. 55, no. 4, pp. 509–522, 2013.
[118] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.,* vol. 32, no. 2, pp. 55–66, Mar. 2015.
[119] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.,* vol. 53, no. 4, pp. 465–494, 2011.
[120] X. Zhu, G. Beauregard, and L. Wyse, "Real-time signal estimation from modified short-time Fourier transform magnitude spectra," *IEEE Trans. Audio Speech Lang. Process.,* vol. 15, no. 5, pp. 1645–1653, July 2007.
[121] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Process. Lett.,* vol. 17, no. 5, pp. 421–424, May 2010.
[122] N. Sturmel and L. Daudet, "Informed source separation using iterative reconstruction," *IEEE Trans. Audio Speech Lang. Process.,* vol. 21, no. 1, pp. 178–185, Jan. 2013.
[123] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.,* vol. 22, no. 12, pp. 1931–1940, Dec. 2014.
[124] P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," *IEEE Signal Process. Lett.,* vol. 20, no. 12, pp. 1235–1239, Dec. 2013.
[125] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.,* vol. 13, no. 5, pp. 1035–1047, Sept. 2005.
[126] A. Sugiyama and R. Miyahara, "Tapping-noise suppression with magnitude-weighted phase-based detection," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.
[127] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE Signal Process. Lett.,* vol. 20, no. 3, pp. 217–220, Mar. 2013.
[128] T. Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase," *IEEE Trans. Signal Process.,* vol. 62, no. 16, pp. 4199–4208, Aug. 2014.
[129] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Process. Lett.,* vol. 20, no. 2, pp. 129–132, Feb. 2013.
[130] M. Parchami, W.-P. Zhu, and B. Champagne, "Microphone array based speech spectral amplitude estimators with phase estimation," in *Proc. IEEE Int. Symp. Circuits and Systems*, 2014, pp. 133–136.
[131] J. Freudenberger, S. Stenzel, and B. Venditti, "A noise PSD and cross-PSD estimation for two-microphone speech enhancement systems," in *Proc. IEEE/SP 15th Workshop on Statistical Signal Processing*, 2009, pp. 709–712.
[132] F. Kallel, M. Ghorbel, M. Frikha, C. Berger-Vachon, and A. B. Hamida, "A noise cross PSD estimator based on improved minimum statistics method for two-microphone speech enhancement dedicated to a bilateral cochlear implant," *Appl. Acoust.,* vol. 73, no. 3, pp. 256–264, 2012.
[133] N. Yousefian and P. Loizou, "A dual-microphone speech enhancement algorithm based on the coherence function," *IEEE Trans. Audio Speech Lang. Process.,* vol. 20, no. 2, pp. 599–609, Feb. 2012.
[134] M. Rahmani, A. Akbari, B. Ayad, and B. Lithgow, "Noise cross PSD estimation using phase information in diffuse noise field," *Signal Process.,* vol. 89, no. 5, pp. 703–709, 2009.
[135] M. Rahmani, A. Akbari, and B. Ayad, "An iterative noise cross-PSD estimation for two-microphone speech enhancement," *Appl. Acoust.,* vol. 70, no. 3, pp. 514–521, 2009.
[136] C. Nelke, C. Beaugeant, and P. Vary, "Dual microphone noise PSD estimation for mobile phones in hands-free position exploiting the coherence and speech presence probability," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2013, pp. 7279–7283.
[137] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 5, pp. 1072–1077, July 2010.
[138] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans Audio Speech Lang Process.*, vol. 19, no. 7, pp. 2159–2169, Sept. 2011.
[139] V. B. Truong, D. M. Nguyen, and Q. H. Dang, "An improved noise reduction algorithm for speech signals using a microphone array," in *Proc. IEEE 5th Int. Conf. Communications and Electronics*, 2014, pp. 472–477.
[140] R. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 1, pp. 223–233, Jan. 2012.
[141] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *Proc. 20th European. Signal Processing Conf.*, 2012, pp. 295–299.
[142] Q. Gong, B. Champagne, and P. Kabal, "Noise power spectral density matrix estimation based on modified IMCRA," in *Proc. 48th Asilomar Conf. Signals, Systems and Computers*, 2014, pp. 1389–1395.
[143] M. Parchami, W.-P. Zhu, and B. Champagne, "A new algorithm for noise PSD matrix estimation in multi-microphone speech enhancement based on recursive smoothing," in *Proc. IEEE Int. Symp. Circuits and Systems*, 2015, pp. 429–432.