Spectral masking and filtering

Timo Gerkmann, Emmanuel Vincent

To cite this version:

Timo Gerkmann, Emmanuel Vincent. Spectral masking and filtering. Emmanuel Vincent; Tuomas Virtanen; Sharon Gannot. Audio source separation and speech enhancement, Wiley, 2018, 978-1-119-27989-1. hal-01881425

HAL Id: hal-01881425

https://hal.inria.fr/hal-01881425

Submitted on 25 Sep 2018

5 Spectral masking and filtering

Timo Gerkmann and Emmanuel Vincent

In this chapter and the following ones, we consider the case of a single-channel input signal ($I = 1$). We denote it as $x(t)$ and omit the channel index $i = 1$ for legibility.

As discussed in Chapter 3, spatial diversity can hardly be exploited to separate such a signal, due to the difficulty of disambiguating the transfer function from the spectrum of each source. Therefore, single-channel separation and enhancement must rely on spectral diversity and exploit properties of the sources such as those listed in Chapter 2. Disregarding phase, one can then separate or enhance the sources using real-valued filters in the time-frequency domain known as *time-frequency masks*.

In the following, we define the concept of time-frequency masking in Section 5.1. We introduce different models to derive a mask from the signal statistics in Section 5.2 and modify them in order to improve perceptual quality in Section 5.3. We summarize the main findings and provide links to forthcoming chapters and more advanced topics in Section 5.4.

5.1 Time-frequency masking

5.1.1 Definition and types of masks

Following the discussion in Chapter 2, filtering is performed in the time-frequency domain (Ephraim and Malah, 1984; Roweis, 2001; Benaroya *et al.*, 2006). Denoting by $x(n, f)$ the complex-valued time-frequency coefficients of the input signal, separation and enhancement can be achieved by

$$\widehat{c}_j(n, f) = w_j(n, f)\, x(n, f) \tag{5.1}$$

or

$$\widehat{s}_j(n, f) = w_j(n, f)\, x(n, f), \tag{5.2}$$

depending whether one wishes to estimate the spatial image $c_j(n, f)$ of source j, or its direct path component that is a delayed and attenuated version of the original source signal $s_j(n, f)$. The filter $w_j(n, f)$ is generally assumed to be real-valued and it is often additionally assumed to satisfy the following constraints for all n, f:

$$0 \leq w_j(n, f) \leq 1 \quad \text{and} \quad \begin{cases} \sum_{j=1}^{J} w_j(n, f) = 1 & \text{in (5.1)} \\ \sum_{j=1}^{J} w_j(n, f) \leq 1 & \text{in (5.2).} \end{cases} \tag{5.3}$$

Such a filter is called a time-frequency mask, a spectral mask, or a masking filter because it operates by selectively hiding unwanted time-frequency areas. The constraints ensure that the sum of the estimated source spatial images $\sum_{j=1}^{J} \widehat{c}_j(n, f)$ is equal to the mixture $x(n, f)$ as per (3.4), and that the sum of the estimated direct path components $\sum_{j=1}^{J} \widehat{s}_j(n, f)$ is smaller than $x(n, f)$ due to the reduction of reverberation.

Masks can be broadly categorized depending on the value range of $w_j(n, f)$. *Binary masks* take binary values $w_j(n, f) \in \{0, 1\}$. They have enjoyed some popularity in the literature due to their ability to effectively improve speech intelligibility in the presence of noise or multiple talkers despite their simplicity (Wang, 2005; Li and Loizou, 2008). *Soft masks* or *ratio masks*, by contrast, can take any value in the range $[0, 1]$.

5.1.2 Oracle mask

In order to understand the potential of time-frequency masking, it is useful to consider the notion of *ideal or oracle mask*, that is the best possible mask for a given signal. This mask can be computed only on development data for which the target signal is known. It provides an upper bound on the separation or enhancement performance achievable.

For most time-frequency representations, the oracle mask cannot easily be computed due to the nonorthogonality of the transform. In practice, this issue is neglected and the oracle mask is computed in each time-frequency bin separately. For the separation of $c_j(n, f)$, for instance, the oracle mask is defined as

$$\widehat{w}_j(n, f) = \underset{w_j(n,f)}{\operatorname{argmin}} \, |c_j(n, f) - w_j(n, f)\, x(n, f)|^2. \tag{5.4}$$

In order to solve this optimization problem under the constraints in (5.3), it is useful to define the real part of the ratio of time-frequency coefficients of the source and the mixture: $r_j(n, f) = \Re(c_j(n, f)/x(n, f))$. In the simplest case when there are only two sources ($J = 2$), the oracle binary masks for the two sources are given by[1)]

$$\widehat{w}_j^{\text{bin}}(n, f) = \begin{cases} 1 & \text{if } r_j(n, f) > \frac{1}{2}, \\ 0 & \text{otherwise,} \end{cases} \tag{5.5}$$

[1)] When $r_j(n, f) = \frac{1}{2}$ for both sources, $\widehat{w}_j^{\text{bin}}(n, f)$ can be arbitrarily set to 1 for either source.

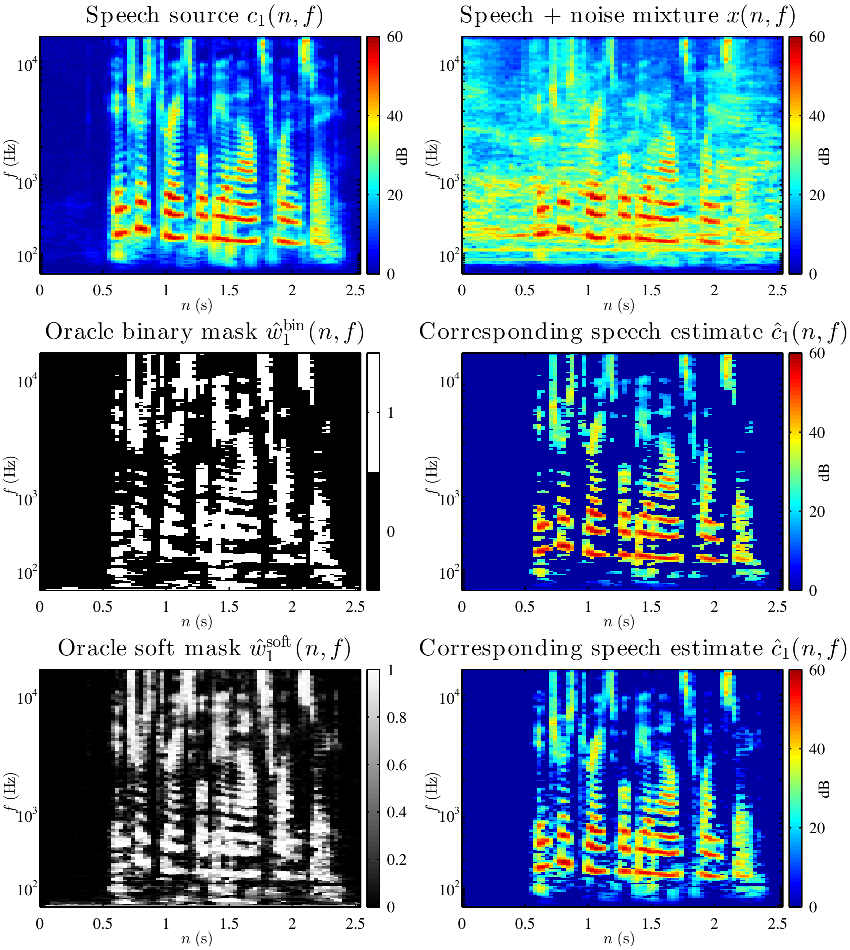

Figure 5.1 Separation of speech from cafe noise by binary vs. soft masking. The masks shown in this example are oracle masks.

and the oracle soft masks by

$$\widehat{w}_j^{\text{soft}}(n,f) = \begin{cases} 0 & \text{if } r_j(n,f) < 0, \\ 1 & \text{if } r_j(n,f) > 1, \\ r_j(n,f) & \text{otherwise.} \end{cases} \tag{5.6}$$

See Vincent *et al.* (2007) for a proof of this result and for the general solution with three or more sources.

These two types of masks are displayed in Fig. 5.1. We see that time-frequency masking can potentially lead to very good separation performance. Also, soft masking appears to perform slightly better than binary masking. As a matter of fact, it has been shown that soft masking improves both speech intelligibility (Jensen and Hendriks, 2012; Madhu *et al.*, 2013) and the maximum achievable signal-to-distortion ratio (SDR) by 3 dB compared to binary masking (Vincent *et al.*, 2007).

5.2 Mask estimation given the signal statistics

In this section we discuss different ways of obtaining filter masks that separate the desired signal from competing sources. We consider the following signal model

$$x(n,f) = c(n,f) + u(n,f) \tag{5.7}$$

where $c(n,f)$ is the target signal and $u(n,f)$ is an uncorrelated interference. For instance, $c(n,f) = c_j(n,f)$ may be the spatial image of the target source and $u(n,f) = \sum_{j' \neq j} c_{j'}(n,f)$ the superposition of all other sources. Alternatively, $c(n,f)$ may be the direct path component of the target source and late reverberation may be modeled as additive, uncorrelated to the target, and comprised in $u(n,f)$ (Lebart *et al.*, 2001). While the assumption that late reverberation is uncorrelated to the target is debatable, it yields powerful and robust estimators in practice (Lebart *et al.*, 2001; Habets, 2007; Cauchi *et al.*, 2015). Then, the derivation of the filters is rather general, meaning that we can use the same spectral mask estimators for signal enhancement, dereverberation, and source separation. The difference in spectral masking based signal enhancement, dereverberation, and source separation rather lies in the way the signals are statistically modeled and how the corresponding parameters, e.g., the power spectra of target and interference are estimated. Due to the fact that spectral masking is applied in each time-frequency bin and for each source independently, we will drop indices j, n, f, in the following unless needed.

5.2.1 Spectral subtraction

Probably the simplest and earliest method for interference reduction is the concept of *spectral subtraction* (Boll, 1979; Berouti *et al.*, 1979). In its simplest form, the average interference magnitude spectrum $\overline{|u|} = \frac{1}{N}\sum_{n=0}^{N-1} |u(n,f)|$ is subtracted from the magnitude spectral coefficients $|x|$ of the mixture and combined with the phase $\angle x$ of the mixture (Boll, 1979):

$$\widehat{c} = (|x| - \overline{|u|})e^{\jmath \angle x}. \tag{5.8}$$

This spectral subtraction rule can be represented by means of a mask w_{SS} as

$$\widehat{c} = \underbrace{\left(1 - \frac{\overline{|u|}}{|x|}\right)}_{w_{\mathrm{SS}}} x = w_{\mathrm{SS}} x. \tag{5.9}$$

(5.8) and (5.9) present the simplest forms of spectral subtraction and are somewhat heuristically motivated. It is important to realize that even though the complex spectral coefficients of the target and the interference are additive, neither their magnitudes nor the averages or the expected values of their magnitudes are additive (with the exception of trivial phases):

$$|c| \neq |x| - |u| \tag{5.10}$$

$$|c| \neq |x| - \overline{|u|} \tag{5.11}$$

$$\mathbb{E}\{|c|\} \neq \mathbb{E}\{|\widehat{c}|\} = \mathbb{E}\{|x|\} - \mathbb{E}\{|u|\}. \tag{5.12}$$

Thus, from a mathematical perspective, the simple spectral amplitude subtraction rule is not optimal.

This is somewhat improved when spectral subtraction is defined on power spectral coefficients, leading to power spectral subtraction:

$$\widehat{c} = (|x|^2 - \overline{|u|^2})^{\frac{1}{2}} e^{\jmath \angle x} \tag{5.13}$$

$$= \underbrace{\left(1 - \frac{\overline{|u|^2}}{|x|^2}\right)^{\frac{1}{2}}}_{w_{\text{PSS}}} x. \tag{5.14}$$

When the temporal average is interpreted as an estimate of the noise power spectrum, i.e., $\overline{|u|^2} = \widehat{\sigma}_u^2$, the quantity $|x|^2 - \overline{|u|^2}$ can be interpreted as an estimate of the target power spectrum $\sigma_c^2 = \mathbb{E}\{|c|^2\}$ (Hendriks *et al.*, 2013). Under the additive signal model (5.7) with zero-mean uncorrelated target and interference, we have $\mathbb{E}\{cx^*\} = 0$ and

$$|c|^2 \neq |x|^2 - \overline{|u|^2}, \text{ but} \tag{5.15}$$

$$\mathbb{E}\{|c|^2\} = \mathbb{E}\{|x|^2\} - \mathbb{E}\{\overline{|u|^2}\} = \mathbb{E}\{|\widehat{c}|^2\}, \tag{5.16}$$

i.e., the power subtraction rule (5.14) represents an unbiased estimator of the target power spectrum $\mathbb{E}\{|c|^2\}$. The zero-mean assumption stems from the fact that the phase is assumed to be uniformly distributed. From a practical viewpoint, the application of power spectral subtraction only requires an estimate of the interference power spectrum $\widehat{\sigma}_u^2 = \overline{|u|^2}$.

5.2.2 Wiener filtering

A more rigorous way of finding a spectral mask w is based on minimizing the mean square error (MSE) between the target c and the estimate $\widehat{c} = w^* x$. Similarly to the problem statement for the oracle mask (5.4), this can be written as

$$w_{\text{SWF}} = \underset{w}{\operatorname{argmin}}\, \mathbb{E}\{|c - w^* x|^2\}. \tag{5.17}$$

The resulting spectral mask is called the *single-channel Wiener filter*. In this expression, both the target spectral coefficients c and the mixture spectral coefficients x are considered as random variables while the spectral mask w is considered to be deterministic and independent of x. In contrast to (5.4), we now consider the possibility that w be complex-valued. Such masking filters are referred to as linearly constrained filters because the estimate $\widehat{c}$ is expressed as a linear function of the mixture x. Thus, the Wiener filter is the linear *minimum mean square error* (MMSE) estimator.

Using the linearity of expectation, we can rephrase the cost in (5.17) as

$$\mathbb{E}\{|c - w^* x|^2\} = \mathbb{E}\{|c|^2\} + |w|^2 \mathbb{E}\{|x|^2\} - 2\Re(w^* \mathbb{E}\{c^* x\}) \tag{5.18}$$

First, let us look at the phase of w. Obviously, the phase of w only influences the last term $-2\Re(w^*\mathbb{E}\{c^*x\})$. Thus, to minimize (5.18) we need to maximize the real part of $w^*\mathbb{E}\{c^*x\}$. It is easy to show that this happens when the phase of w is the same as that of $\mathbb{E}\{c^*x\}$ such that the product $w^*\mathbb{E}\{c^*x\}$ is real-valued:

$$\angle w_{\mathrm{SWF}} = \angle \mathbb{E}\{c^* x\}. \tag{5.19}$$

Secondly, let us look at the magnitude of w. The optimal magnitude can be found by inserting the optimal phase (5.19) and equating the derivative with respect to $|w|$ to zero

$$\begin{aligned} 0 &= \frac{\partial}{\partial |w|}\left(\mathbb{E}\{|c|^2\} + |w|^2\mathbb{E}\{|x|^2\} - 2|w||\mathbb{E}\{c^*x\}|\right) && (5.20)\\ &= 2|w|\mathbb{E}\{|x|^2\} - 2|\mathbb{E}\{c^*x\}|. && (5.21)\end{aligned}$$

Solving for $|w|$, we obtain

$$|w_{\mathrm{SWF}}| = \frac{|\mathbb{E}\{c^*x\}|}{\mathbb{E}\{|x|^2\}}. \tag{5.22}$$

Combining the optimal magnitude (5.22) and the optimal phase (5.19), we obtain the optimal masking filter

$$w_{\mathrm{SWF}} = \frac{\mathbb{E}\{c^*x\}}{\mathbb{E}\{|x|^2\}}. \tag{5.23}$$

With the assumption that the target c and the interference u are zero-mean and mutually uncorrelated, we have $\mathbb{E}\{c^*x\} = \mathbb{E}\{|c|^2\}$ and $\mathbb{E}\{|x|^2\} = \mathbb{E}\{|c|^2\} + \mathbb{E}\{|u|^2\}$ such that the spectral mask is given by

$$w_{\mathrm{SWF}} = \frac{\mathbb{E}\{|c|^2\}}{\mathbb{E}\{|c|^2\} + \mathbb{E}\{|u|^2\}} = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_u^2}. \tag{5.24}$$

The Wiener filter w_{SWF} turns out to be real-valued and to satisfy the constraints in (5.3). From a practical viewpoint, in contrast to power spectral subtraction, estimates of both the target and the interference power spectra are needed.

Note that, depending on the problem formulation, the Wiener filter can also be complex-valued. For instance, if we aimed at finding the source signal s instead of its spatial image c, we would replace c by s in (5.17). Assuming a narrowband model $x = as + u$, the Wiener filter solution (5.23) would result in

$$w_{\text{SWF}} = \frac{a\sigma_s^2}{|a|^2\sigma_s^2 + \sigma_u^2}. \tag{5.25}$$

As explained in Chapter 2, for commonly chosen frame sizes, the narrowband model does not properly account for late reverberation. An alternative way to perform dereverberation is to model late reverberation as additive and uncorrelated with the target. Then, the late reverberant power spectrum can be incorporated into the interference power spectrum σ_u^2 in (5.24). This approach can yield robust results in practice (Lebart *et al.*, 2001; Habets, 2007; Cauchi *et al.*, 2015).

5.2.3 Bayesian estimation of Gaussian spectral coefficients

In the previous section we showed that if we constrain the estimate to be a linear function of the mixture, the MMSE estimator is given by the Wiener filter (5.24). For its derivation, we did not assume any underlying distribution of the target or interference spectral magnitude coefficients, but only that the target and interference spectral coefficients are zero-mean and mutually uncorrelated. Hence, this result is very compact and general. However, the question arises if we can get an even better result if we allow for a nonlinear relationship between the input x and the output $\widehat{c}(x)$ of the filter. For this, we need to optimize the more general problem

$$\widehat{c}^{\text{Bayes}} = \underset{\widehat{c}}{\operatorname{argmin}}\, \mathbb{E}\{|c - \widehat{c}(x)|^2\} \tag{5.26}$$

where the estimate $\widehat{c}(x)$ is any, possibly nonlinear, function of the mixture x.

It can be shown that solving this general MMSE problem is equivalent to finding the *posterior mean* (Schreier and Scharf, 2010, sec. 5.2), i.e.,

$$\widehat{c}^{\text{Bayes}} = \mathbb{E}\{c \mid x\} = \int c\, p(c \mid x)\mathrm{d}c. \tag{5.27}$$

Thus, the MMSE estimator is also referred to as the posterior mean estimator. The formulation as a posterior mean estimator in (5.27) allows us to elegantly use the concepts of *Bayesian statistics* to find the unconstrained MMSE estimator, i.e., to solve (5.26).

To find the posterior mean (5.27), we need a model for the conditional probability distribution of the searched quantity c, referred to as *posterior* in Bayesian estimation. While finding a model for the posterior is often difficult, using Bayes' theorem the posterior $p(c \mid x)$ can be expressed as a function of the *likelihood* $p(x \mid c)$ and the *prior* $p(c)$ as

$$p(c \mid x) = \frac{p(c, x)}{p(x)} = \frac{p(c, x)}{\int p(c, x)\mathrm{d}c} = \frac{p(c)p(x \mid c)}{\int p(c)p(x \mid c)\mathrm{d}c}. \tag{5.28}$$

This means that instead of the posterior, we now need models for the likelihood and the prior over c in order to solve (5.27).

If we have an additive signal model then, as the target signal is given, the randomness in the likelihood $p(x \mid c)$ is given only by the interference signal. A common assumption is that the interference signal is zero-mean *complex Gaussian* distributed with variance $\sigma_u^2 = \mathbb{E}\{|u|^2\}$. As a consequence, the likelihood is complex Gaussian with mean c and variance σ_u^2:

$$p(x \mid c) = \frac{1}{\pi \sigma_u^2} \exp\left(-\frac{|x-c|^2}{\sigma_u^2}\right). \tag{5.29}$$

This Gaussian interference model is the most popular. Other interference models, e.g., Laplacian, have also been discussed in the literature (Martin, 2005; Benaroya *et al.*, 2006).

While the likelihood requires defining a statistical model for the interference coefficients, the prior corresponds to a statistical model of the target spectral coefficients. The simplest assumption is a zero-mean complex Gaussian model with variance σ_c^2:

$$p(c) = \frac{1}{\pi \sigma_c^2} \exp\left(-\frac{|c|^2}{\sigma_c^2}\right). \tag{5.30}$$

For uncorrelated zero-mean Gaussian target and interference, the sum $x = c + u$ is zero-mean complex Gaussian with variance $\sigma_c^2 + \sigma_u^2$, i.e., the resulting *evidence* model is

$$p(x) = \frac{1}{\pi\left(\sigma_c^2 + \sigma_u^2\right)} \exp\left(-\frac{|x|^2}{\sigma_c^2 + \sigma_u^2}\right). \tag{5.31}$$

Using the Gaussian likelihood (5.29) and prior (5.30) in the numerator and the evidence (5.31) in the denominator of (5.28), we obtain

$$p(c \mid x) = \frac{\frac{1}{\pi\sigma_c^2} \exp\left(-\frac{|c|^2}{\sigma_c^2}\right) \frac{1}{\pi\sigma_u^2} \exp\left(-\frac{|x-c|^2}{\sigma_u^2}\right)}{\frac{1}{\pi(\sigma_c^2+\sigma_u^2)} \exp\left(-\frac{|x|^2}{\sigma_c^2+\sigma_u^2}\right)} \tag{5.32}$$

$$= \frac{1}{\pi \frac{\sigma_c^2\sigma_u^2}{\sigma_c^2+\sigma_u^2}} \exp\left(-\frac{\frac{\sigma_c^2}{\sigma_c^2+\sigma_u^2}|x-c|^2 + \frac{\sigma_u^2}{\sigma_c^2+\sigma_u^2}|c|^2 - \frac{\sigma_c^2\sigma_u^2}{(\sigma_c^2+\sigma_u^2)^2}|x|^2}{\frac{\sigma_c^2\sigma_u^2}{\sigma_c^2+\sigma_u^2}}\right). \tag{5.33}$$

Substituting $\lambda = \frac{\sigma_c^2\sigma_u^2}{\sigma_c^2+\sigma_u^2}$, using $|x-c|^2 = |x|^2 + |c|^2 - 2\Re(xc^*)$ we obtain

$$p(c \mid x) = \frac{1}{\pi\lambda} \exp\left(-\frac{|c|^2 + |x|^2\left(\frac{\sigma_c^2}{\sigma_c^2+\sigma_u^2}\right)^2 - 2\Re(xc^*)\frac{\sigma_c^2}{\sigma_c^2+\sigma_u^2}}{\lambda}\right) \tag{5.34}$$

$$= \frac{1}{\pi\lambda} \exp\left(-\frac{\left|c - \frac{\sigma_c^2}{\sigma_c^2+\sigma_u^2}x\right|^2}{\lambda}\right). \tag{5.35}$$

$p(\theta \mid x)$	Posterior
$p(x \mid \theta)$	Likelihood
$p(\theta)$	Prior
$p(x)$	Evidence

Table 5.1 Bayesian probability distributions for the observation x and the searched quantity θ.

This result is interesting in many ways. First of all, we see that for a Gaussian likelihood and prior, the posterior is also Gaussian. Secondly, we may directly see that the mean of the posterior is given by

$$\mathbb{E}\{c \mid x\} = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_u^2} x = w_{\text{SWF}}\, x = \widehat{c}, \tag{5.36}$$

which is identical to the Wiener solution found as the MMSE linearly constrained filter (5.24). In other words, for Gaussian target and interference, the optimal estimate in the MMSE sense is the Wiener filter no matter whether we constrain the filter to be linear or not. Note, however, that for nongaussian target or interference this is not necessarily true and the MMSE estimate is generally a nonlinear function of the observation. Finally, we also see that the posterior exhibits the variance $\lambda = \frac{\sigma_c^2 \sigma_u^2}{\sigma_c^2 + \sigma_u^2}$ which can also be seen as a measure of uncertainty for the Wiener estimate.

The concept of Bayesian statistics and estimation is a general and powerful tool. In Table 5.1 an overview of the probability distributions relating to the searched quantity θ (so far we considered $\theta = c$) and the observation x is given. Based on these conditional distributions, other estimators can be defined, too. For instance the θ that maximizes the likelihood is referred to as the maximum likelihood (ML) estimate, while the θ that maximizes the posterior is referred to as the *maximum a posteriori* (MAP) estimate (see Table 5.2). Using Bayes' theorem, the MAP estimator can be expressed as a function of the likelihood and the prior as

$$\begin{aligned} \widehat{\theta}^{\text{MAP}} &= \underset{\theta}{\operatorname{argmax}}\, p(\theta \mid x) & (5.37)\\ &= \underset{\theta}{\operatorname{argmax}}\, \frac{p(x \mid \theta)\, p(\theta)}{p(x)} & (5.38)\\ &= \underset{\theta}{\operatorname{argmax}}\, p(x \mid \theta)\, p(\theta) & (5.39) \end{aligned}$$

Here the normalization by the evidence model $p(x)$ is not necessary as it is not a function of the searched quantity θ. Hence, whenever prior information about the searched quantity is given, it can be used to extend the likelihood and to obtain an improved estimator by means of a MAP estimator or, using (5.28), the MMSE estimator.

If the respective conditional distributions are known and unimodal, the ML and MAP estimators defined in Table 5.2 can be obtained by equating the derivative of the conditional distributions with respect to the searched quantity θ to zero and solving

$\widehat{\theta}^{\mathrm{ML}} = \underset{\theta}{\operatorname{argmax}}\, p(x \mid \theta)$	ML estimator
$\widehat{\theta}^{\mathrm{MAP}} = \underset{\theta}{\operatorname{argmax}}\, p(\theta \mid x)$	MAP estimator
$\widehat{\theta}^{\mathrm{MMSE}} = \mathbb{E}\{\theta \mid x\}$	MMSE estimator

Table 5.2 Criteria for the estimation of θ.

for θ. In the Gaussian case considered so far, finding the ML and MAP estimates of c is rather simple. The likelihood (5.29) is maximum when the argument of its exponential function is as small as possible, i.e., when $c = x$. In other words, the ML estimate is equal to the mixture spectral coefficient

$$\widehat{c}^{\mathrm{ML}} = x = w_{\mathrm{ML}} x. \tag{5.40}$$

The resulting masking filter is $w_{\mathrm{ML}} = 1$ for all time-frequency bins and thus does not result in any interference reduction.

As the Gaussian distribution is not only unimodal but also symmetric, the mean of the posterior is identical to its mode, meaning that for any unimodal and symmetric posterior, the MAP estimator is equivalent to the MMSE estimator, i.e., for a Gaussian target and interference model the estimate is given by

$$\widehat{c}^{\mathrm{MAP}} = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_u^2}\, x = w_{\mathrm{MAP}}\, x. \tag{5.41}$$

The MAP solution is thus identical to the MMSE solution, and $w_{\mathrm{MAP}} = w_{\mathrm{SWF}}$.

From the above results, it may seem as if the Bayesian theory is not of much help, as under a Gaussian signal model either a trivial filter arises as the ML estimator in (5.40), or simply the Wiener masking filter arises as the unconstrained MMSE estimator in (5.36) or the MAP estimator in (5.41). However, taking the linear approach (5.17) the Wiener solution is found without any assumptions on the distribution. So why would the Bayesian concept be of importance? The answer is simple: for many random variables the Gaussian assumption is either invalid or impossible to verify and alternative distributions can be assumed. Bayesian estimation then provides a very general concept to find optimal estimators for these nongaussian quantities that may outperform the simple Wiener filter.

5.2.4 Estimation of magnitude spectral coefficients

A prominent example where nongaussianity matters is the estimation of nonnegative quantities such as magnitude or power spectral coefficients. For instance, estimating spectral magnitudes rather than complex spectral coefficients is thought to be perceptually more meaningful (Ephraim and Malah, 1984, 1985) (see also Section 5.3). We now argue that for such nonnegative quantities the Wiener filter is not the optimal solution, neither in the linearly constrained MMSE sense, nor in the Bayesian sense. For this, it is important to note that nonnegative quantities are not zero-mean and hence a multiplicative linear filter as defined in (5.17) may result in a biased estimate, i.e., $\mathbb{E}\{\widehat{\theta}\} \neq \mathbb{E}\{\theta\}$ where θ is the nonnegative searched quantity. To find an unbiased linear estimator that minimizes the MSE, the problem statement in (5.17) must be extended by adding a term m such that the MSE and the bias can both be controlled. Let r be a nonzero-mean mixture and θ be the nonzero-mean searched quantity. Without loss of generality $r = |x|$ could be the amplitude of the mixture while $\theta = |c|$ could be the magnitude of the target spectral coefficients. The problem then becomes

$$\min_{w,m} \mathbb{E}\{|\theta - (wr + m)|^2\}. \tag{5.42}$$

Optimizing for both w and m, the resulting estimator boils down to subtracting the mean of the nonnegative mixture $r' = r - \mathbb{E}\{r\}$ and the nonnegative searched quantity $\theta' = \theta - \mathbb{E}\{\theta\}$, applying a filter similar to (5.23), and adding the mean back of the desired quantity $\mathbb{E}\{\theta\}$ as

$$\widehat{\theta} = \frac{\mathbb{E}\{\theta' r'\}}{\mathbb{E}\{r'^2\}} r' + \mathbb{E}\{\theta\}. \tag{5.43}$$

However, in contrast to (5.24), just as for Bayesian estimators, the relationship between $\mathbb{E}\{\theta r\}$, $\mathbb{E}\{r^2\}$, and $\mathbb{E}\{\theta\}$ depends on the distributions of the random variables and not only on their variances (Hendriks *et al.*, 2013). Therefore, we now put aside linearly constrained estimators and focus on the potentially stronger unconstrained Bayesian estimators.

Besides the insight that nonnegative quantities are not zero-mean, it is also important to realize that they are not Gaussian. This is intuitively clear when considering that Gaussian distributed random variables take values between $-\infty$ and $+\infty$ per definition, while nonnegative random variables can only take values between 0 and $+\infty$ per definition. As a result, the Wiener filter (5.24) obtained as the MMSE and MAP solutions (5.41) under a Gaussian signal model is not the optimal estimator for nonnegative quantities.

As a concrete example, we consider the Bayesian estimation of spectral magnitudes, i.e., $\theta = |c|$, obtained from complex Gaussian distributed spectral coefficients c following (5.30). The prior $p(\theta) = p(|c|)$ can be obtained by transforming $p(c)$ (5.30) into polar coordinates and integrating out the phase. The result is the well known *Rayleigh* distribution

$$p(|c|) = \frac{2|c|}{\sigma_c^2} \exp\left(-\frac{|c|^2}{\sigma_c^2}\right). \tag{5.44}$$

The posterior could then be obtained from this prior and the likelihood (5.29) using Bayes theorem and integration as in (5.28). Alternatively, the posterior $p(\theta \mid x) = p(|c| \mid x)$ can be obtained by directly transforming $p(c \mid x)$ (5.35) into polar coordinates and integrating out the phase. It is well known that the resulting magnitude posterior follows a *Rician* distribution (Wolfe and Godsill, 2003)

$$p(|c| \mid x) = \frac{2|c|}{\lambda} \exp\left(-\frac{|c|^2 + w_{\text{SWF}}^2|x|^2}{\lambda}\right) I_0\left(\frac{2|x|\,|c|\,w_{\text{SWF}}}{\lambda}\right), \tag{5.45}$$

with λ defined below (5.33), w_{SWF} defined in (5.36), and $I_0(\cdot)$ the modified zeroth-order Bessel function of the first kind (Gradshteyn and Ryzhik, 2000). The MMSE optimal estimator of spectral magnitudes is thus given as the mean of the Rician distribution (5.45). Just as for other common distributions, the mean of the Rician distribution can nowadays easily be found in the literature. In the context of speech enhancement the resulting estimator has been proposed by Ephraim and Malah (1984) and referred to as the *short-time spectral amplitude* estimator:

$$\mathbb{E}\{|c| \mid x\} = \Gamma(1.5)\,\sqrt{\lambda}\,\Phi\left(-0.5, 1; -w_{\text{SWF}}\frac{|x|^2}{\sigma_u^2}\right), \tag{5.46}$$

where $\Gamma(\cdot)$ is the gamma function (Gradshteyn and Ryzhik, 2000, (8.31)), $\Gamma(1.5) = \sqrt{\pi}/2$, and $\Phi(\cdot,\cdot;\cdot)$ is the confluent hypergeometric function (Gradshteyn and Ryzhik, 2000, (9.210)). An estimate of the target complex coefficients is then obtained by combining the amplitude estimate (5.46) with the phase of the mixture. A spectral masking filter can be obtained by dividing the estimated target coefficients by the mixture coefficients. However, the relationship between the mixture and the estimate remains nonlinear and the resulting spectral masking filter may not satisfy the constraints in (5.3).

Furthermore, as the Rician distribution is not symmetric, the mode and the mean of the posterior are not identical anymore. Thus, in contrast to the Gaussian case discussed in Section 5.2.3, the MAP estimator of spectral magnitudes is different from the MMSE estimate (5.46). This is also illustrated in Fig. 5.2 where the Rician posterior (5.45) is shown along with the MMSE and the MAP estimates. While analytically finding the mode of the Rician distribution is difficult, Wolfe and Godsill (2003) proposed an approximate closed-form solution for the MAP estimator of target magnitudes which is easier to implement than the MMSE estimator in (5.46):

$$|c|^{\text{MAP}} \approx \left(\frac{1}{2}w_{\text{SWF}} + \sqrt{\left(\frac{1}{2}w_{\text{SWF}}\right)^2 + \frac{\lambda}{4|x|^2}}\right)|x| \tag{5.47}$$

The larger the argument of the Bessel function in (5.45), the better this approximation holds. In Fig. 5.2 an example for the approximate MAP estimate is given.

5.2.5 Heavy-tailed priors

Another situation when Bayesian estimation helps is when considering nongaussian distributions for the source spectral coefficients. While we previously argued that the Gaussian model is a very useful and generic model, alternative distributions can also be assumed (Martin, 2005). Indeed, it is impossible to know the true distribution of the source spectral coefficients due to the fact that it is nonstationary, i.e., it varies from one time-frequency bin to another, and it cannot be estimated from the observation in a single time-frequency bin. One attempt to estimate the distribution of clean speech is to compute the histogram for a narrow range of estimated speech powers $\widehat{\sigma}_c^2$

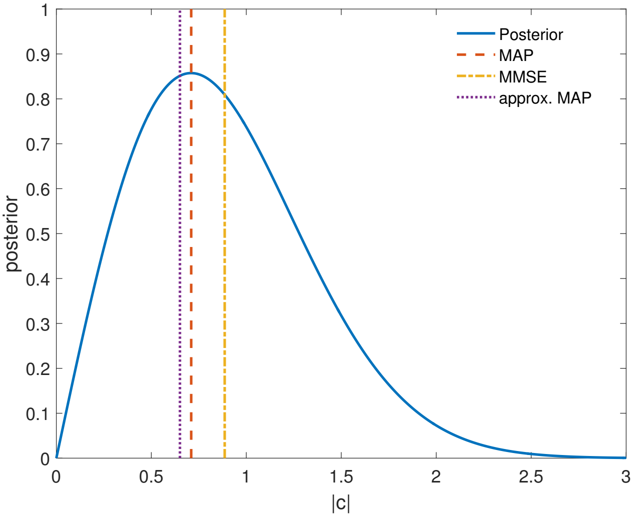

Figure 5.2 Illustration of the Rician posterior $p(|c| \mid x)$ (5.45) for $\sigma_c^2 = 1$, $\sigma_u^2 = 10$, and $|x| = \sqrt{11}$. The red dashed line shows the mode of the posterior and thus the MAP estimate of target spectral magnitudes $|c|$. The purple dotted line corresponds to the approximate MAP estimate (5.47), and the yellow dash-dotted line corresponds to the posterior mean (5.46) and thus the MMSE estimate of $|c|$.

(Martin, 2005). A second approach is to normalize the speech spectral coefficients by the square-root of the estimated speech power $\sqrt{\widehat{\sigma_c^2}}$ and to compute the histogram over time-frequency bins where speech is active (Gerkmann and Martin, 2010). In both cases, the found distributions also depend on the chosen spectral transformation, frame size and power spectrum estimator (Gerkmann and Martin, 2010). However, whatever the choices made, the obtained histogram typically follows a *heavy-tailed* distribution, also known as a *supergaussian* or *sparse* distribution. This means that small and large values are more likely and medium values are less likely compared to a Gaussian distribution. In Fig. 5.3, an example of the histogram of normalized speech coefficients is shown (Gerkmann and Martin, 2010). Here, the histogram of the real part of complex speech coefficients $\Re(c)$ is compared to a Gaussian and a *Laplacian* distribution. Clearly, the histogram is more similar to the heavy-tailed Laplacian distribution. We emphasize that the Laplacian distribution is only an example of a heavy-tailed distribution and sparser distributions may often provide a better fit (Vincent, 2007).

If the target spectral coefficients follow a heavy-tailed distribution, nonlinear estimators can be derived that outperform the simple Wiener filter. This idea has already been discussed by Porter and Boll (1984), who proposed to obtain the optimal filter based on training data. Martin proposed a closed-form solution when the target follows a gamma (Martin, 2002) or Laplacian (Martin and Breithaupt, 2003) prior distribution and showed that improved performance can be achieved.

As a consequence, in the last decade many more proposals and improvements us-

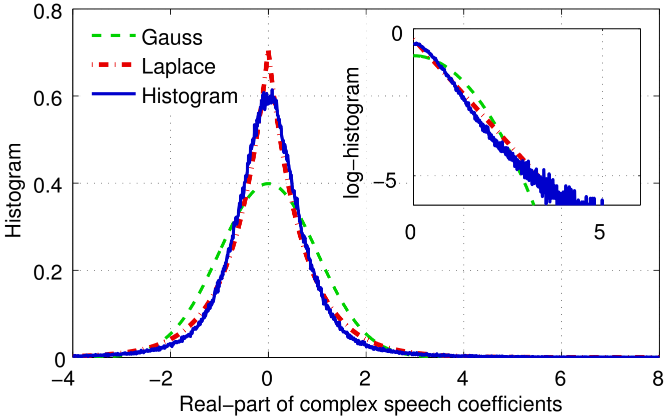

Figure 5.3 Histogram of the real part of complex speech coefficients (Gerkmann and Martin, 2010).

ing heavy-tailed priors were proposed. For this, parameterizable speech priors were proposed in order to optimize the achieved results by means of instrumental measures or listening experiments. One example of such a parameterizable speech prior is the χ *distribution* for spectral amplitudes $|c|$ as

$$p(|c|) = \frac{2}{\Gamma(\mu)} \left(\frac{\mu}{\sigma_c^2}\right)^{\mu} |c|^{2\mu-1} \exp\left(-\frac{\mu}{\sigma_c^2}|c|^2\right) \tag{5.48}$$

with $\Gamma(\cdot)$ as defined below (5.46). The so-called shape parameter μ controls the heavy-tailedness of the target prior. While for $\mu = 1$ (5.48) corresponds to the Rayleigh distribution (5.44) thus implying complex Gaussian coefficients c, for $0 < \mu < 1$ a heavy-tailed speech prior results. Using different heavy-tailed priors, both MAP and MMSE estimators were derived for complex spectral coefficients, spectral amplitudes, and compressed spectral amplitudes (Martin, 2005; Lotter and Vary, 2005; Benaroya *et al.*, 2006; Erkelens *et al.*, 2007; Breithaupt *et al.*, 2008b). In Fig. 5.4 it can be seen that for a large input $x/\sigma_c \gg 1$, a heavy-tailed prior results in a larger output than when a Gaussian prior is used. This is because using heavy-tailed target priors, outliers in the input are more likely attributed to the target signal. This behavior of supergaussian estimators results in less speech attenuation and thus less target distortion in the processed signal. However, this behavior may also increase the amount of undesired outliers in the processed signal that may be perceived as annoying musical tones.

5.2.6 Masks based on source presence statistics

While in deriving linear MMSE estimators via (5.17) a spectral masking filter was explicitly estimated, for the Bayesian estimators considered so far, the targets were

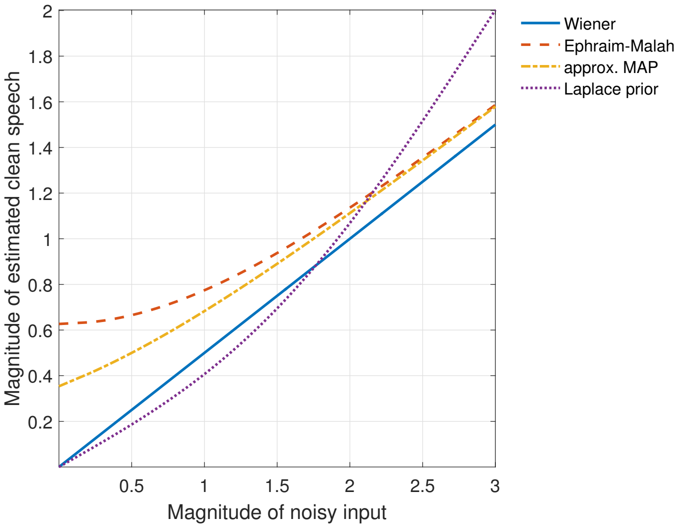

Figure 5.4 Input-output characteristics of different spectral filtering masks. In this example $\sigma_c^2 = \sigma_u^2 = 1$. “Wiener” refers to the Wiener filter, “Ephraim-Malah” to the short-time spectral amplitude estimator of Ephraim and Malah (1984), and “approx. MAP” to the approximate MAP amplitude estimator (5.47) of Wolfe and Godsill (2003). While “Wiener”, “Ephraim-Malah”, and “approx. MAP” are based on a Gaussian speech model, “Laplace prior” refers to an estimator of complex speech coefficients with a supergaussian speech prior (Martin and Breithaupt, 2003). Compared to the linear Wiener filter, amplitude estimators tend to apply less attenuation for low inputs, while supergaussian estimators tend to apply less attenuation for high inputs.

(functions of) the target coefficients c. A spectral masking filter could then be obtained by dividing input and output as $w = \widehat{c}/x$.

A different way to obtain a spectral masking filter using Bayesian statistics is to estimate the source presence probability. For this, we define $\mathcal{H}_1$ as the hypothesis that the target is active in the considered time-frequency bin, while $\mathcal{H}_0$ denotes the hypothesis that it is inactive. A spectral masking filter can then be obtained by computing the posterior probability that the target is present. Using Bayes’ theorem, similar to (5.28), this posterior probability can be expressed as

$$w = P(\mathcal{H}_1 \mid x) = \frac{p(x \mid \mathcal{H}_1)P(\mathcal{H}_1)}{p(x \mid \mathcal{H}_1)P(\mathcal{H}_1) + p(x \mid \mathcal{H}_0)P(\mathcal{H}_0)}. \tag{5.49}$$

As opposed to the previously described mask estimators, the optimality criterion is not on the separation or enhancement of sources anymore but merely on the estimation of their presence. The source presence probability is a particular powerful tool, e.g., to estimate parameters such as the interference power spectrum. More details about speech presence probability (SPP) estimation can be found in Chapter 6.

5.3 Perceptual improvements

Whenever applying a spectral mask, undesired artifacts may occur. When the mask in a given time-frequency bin is lower than it should, target distortion occurs. When the mask is larger, interference reduction is limited. The artifacts are even more disturbing when the mask exhibits large values in isolated time-frequency regions that result in isolated outliers in the estimated signal. In the time domain, these isolated spectral peaks result in sinusoidal components of short duration and are often perceived as annoying musical noise. An early proposal to reduce musical noise artifacts is to *overestimate* the interference power spectrum in order to reduce spectral outliers in the mask (Berouti *et al.*, 1979) at the costs of increased target distortion. Another way is to apply a lower limit, so-called *spectral floor* (Berouti *et al.*, 1979) to the spectral masking filter at the cost of lesser interference reduction. Using this spectral floor, both musical noise and target distortion are controlled and perceptually convincing results can be obtained.

Porter and Boll (1984) observed that when spectral masking filters are derived by estimating compressed spectral amplitudes, musical noise can be reduced. As a compression of amplitudes is related to the way we perceive the loudness of sounds, also the estimation of compressed spectral coefficients is considered to be perceptually more meaningful than an amplitude estimation without compression. Ephraim and Malah (1985) were the first to derive a closed-form solution for a Bayesian estimator of logarithmically compressed amplitudes under a Gaussian prior and likelihood. You *et al.* (2005) derived a more general estimator for powers of spectral amplitudes as $\mathbb{E}\{|c|^\beta \mid x\}$ that also generalizes the square root ($\beta = 1/2$) and logarithmic ($\beta \to 0$) compression. Breithaupt *et al.* (2008b) again generalized this result for the parameterizable prior (5.48), thus enabling the estimation of compressed spectral coefficients under heavy-tailed priors. This flexible estimator results in

$$\mathbb{E}\{|c|^\beta \mid x\} = \left(\frac{\sigma_c^2\sigma_u^2}{\sigma_c^2 + \mu\sigma_u^2}\right)^{\frac{\beta}{2}} \frac{\Gamma(\mu + \beta/2)}{\Gamma(\mu)} \frac{\Phi(1-\mu-\beta/2, 1; -\nu)}{\Phi(1-\mu, 1; -\nu)} \tag{5.50}$$

with $\nu = |x|^2\sigma_c^2/(\sigma_u^4\mu + \sigma_u^2\sigma_c^2)$ and $\Gamma(\cdot)$, $\Phi(\cdot)$ as defined below (5.46). Compression is obtained for $0 < \beta < 1$.

Another way to reduce processing artifacts is to apply smoothing methods to the spectral masking filter or its parameters (Vincent, 2010). This has to be done with great care in order not to introduce smearing artifacts. In the single-channel case, simple nonadaptive temporal smoothing often does not lead to satisfactory results and adaptive smoothing methods are used instead (Ephraim and Malah, 1984; Cappé, 1994; Martin and Lotter, 2001). Good results can also be achieved by carefully smoothing over both time and frequency (Cohen and Berdugo, 2001; Gerkmann *et al.*, 2008), or smoothing in perceptually motivated filter bands (Esch and Vary, 2009; Brandt and Bitzer, 2009). An elegant way to incorporate typical speech spectral structures in the smoothing process is to apply so-called *cepstral smoothing* (Breithaupt *et al.*, 2007), as illustrated in Fig. 5.5. The *cepstrum* is defined as the spectral transform of the logarithmic amplitude spectrum. In this domain speech-like spectral structures are compactly represented by few lower cepstral coefficients that represent the speech spectral envelope, and a peak in the upper cepstrum that represents the spectral fine structure of voiced speech. Thus, in the cepstral domain the speech related coefficients can be preserved while smoothing is mainly applied to the remaining coefficients that represent spectral structures that are not speech-like. This method can be applied directly to spectral masks (Breithaupt *et al.*, 2007), or to the target and interference spectra from which the masks are computed, e.g., (Breithaupt *et al.*, 2008a; Gerkmann *et al.*, 2010).

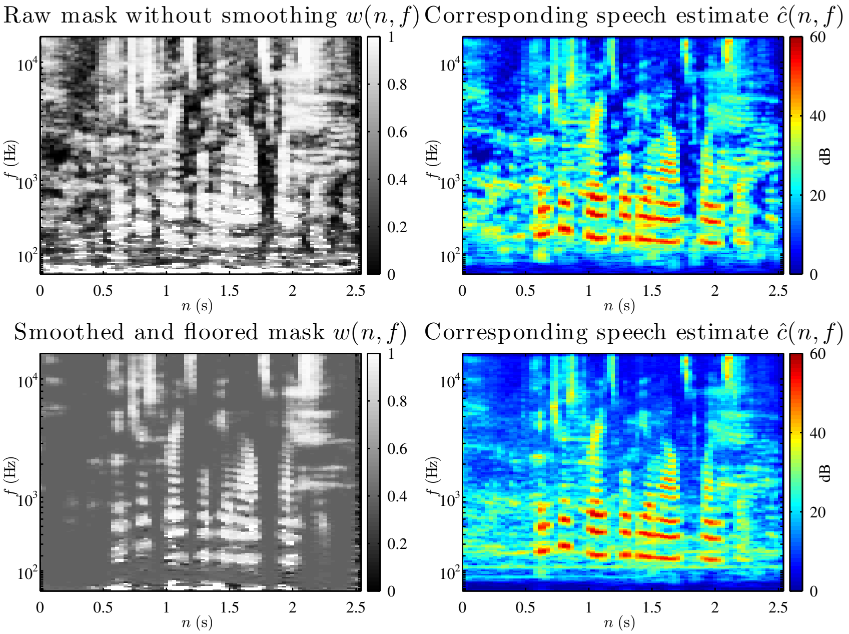

Figure 5.5 Examples of estimated filters for the noisy speech signal in Fig. 5.1. The filters were computed in the STFT domain but are displayed on a nonlinear frequency scale for visualization purposes.

5.4 Summary

In this chapter, we introduced the concept of spectral masking for signal enhancement and separation. We reviewed different ways of deriving time-frequency masks, from spectral subtraction and Wiener filtering to more general Bayesian estimation of source spectral coefficients or activity masks. We argued that when functions of complex coefficients, such as magnitudes, are targeted better estimators than spectral

Method	Pros	Cons
Spectral subtraction	Simple	Somewhat heuristic
Bayesian estimation	Very flexible, well defined optimality criteria	Models needed, closed-form solutions not guaranteed to exist
Wiener filtering	Simple	Not optimal for nonnegative or nongaussian quantities
Heavy-tailed priors	Less target distortion	Often more musical noise
Source presence statistics	Powerful tool to estimate parameters such as the interference power spectrum	Optimality not defined in terms of the separated/enhanced signals
Perceptual improvements	Better sound quality	Often more interference

Table 5.3 Overview of the discussed estimation schemes.

subtraction and Wiener filtering can be derived using Bayesian estimation. We finally discussed methods to control the perceptual quality of the output by heuristic tweaks, estimation of compressed spectral coefficients, and time-frequency smoothing.

The reviewed estimators depend on the statistics of the source signals, namely the source activities $P(\mathcal{H}_1 \mid x)$ and $P(\mathcal{H}_0 \mid x)$ (zeroth-order statistics) or the source variances σ_c^2 and σ_u^2 (second-order statistics). In the case of three or more sources, these boil down to estimation the activities or the variances $\sigma_{c_j}^2$ of all sources. The estimation of these quantities in the single-channel case is covered in Chapters 6, 7, 8, and 9. An overview of the discussed mask estimators is given in Table 5.3.

In recent years, improved estimators that go beyond the time-frequency masking paradigm have been proposed. For instance, researchers showed that the correlation neighboring time-frequency bins and the spectral phase can be estimated and exploited for enhancement with reduced distortion. These techniques are reviewed in Chapter 19.

Bibliography

Benaroya, L., Bimbot, F., and Gribonval, R. (2006) Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech, and Language Processing*, **14** (1), 191–199.

Berouti, M., Schwartz, R., and Makhoul, J. (1979) Enhancement of speech corrupted by acoustic noise, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 208–211.

Boll, S.F. (1979) Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **27** (2), 113–120.

Brandt, M. and Bitzer, J. (2009) Optimal spectral smoothing in short-time spectral attenuation (STSA) algorithms: Results of objective measures and listening tests, in *Proceedings of European Signal Processing Conference*, pp. 199–203.

Breithaupt, C., Gerkmann, T., and Martin, R. (2007) Cepstral smoothing of spectral filter gains for speech enhancement without musical noise. *IEEE Signal Processing Letters*, **14** (12), 1036–1039.

Breithaupt, C., Gerkmann, T., and Martin, R. (2008a) A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp.

4897–4900.

Breithaupt, C., Krawczyk, M., and Martin, R. (2008b) Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 4037–4040.

Cappé, O. (1994) Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, **2** (2), 345–349.

Cauchi, B., Kodrasi, I., Rehr, R., Gerlach, S., Jukic, A., Gerkmann, T., Doclo, S., and Goetze, S. (2015) Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech. *EURASIP Journal on Advances in Signal Processing*, **2015** (61), 1–12.

Cohen, I. and Berdugo, B. (2001) Speech enhancement for non-stationary noise environments. *Signal Processing*, **81** (11), 2403–2418.

Ephraim, Y. and Malah, D. (1984) Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **32** (6), 1109–1121.

Ephraim, Y. and Malah, D. (1985) Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **33** (2), 443–445.

Erkelens, J.S., Hendriks, R.C., Heusdens, R., and Jensen, J. (2007) Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors. *IEEE Transactions on Audio, Speech, and Language Processing*, **15** (6), 1741–1752.

Esch, T. and Vary, P. (2009) Efficient musical noise suppression for speech enhancement systems, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 4409–4412.

Gerkmann, T., Breithaupt, C., and Martin, R. (2008) Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors. *IEEE Transactions on Audio, Speech, and Language Processing*, **16** (5), 910–919.

Gerkmann, T., Krawczyk, M., and Martin, R. (2010) Speech presence probability estimation based on temporal cepstrum smoothing, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 4254–4257.

Gerkmann, T. and Martin, R. (2010) Empirical distributions of DFT-domain speech coefficients based on estimated speech variances, in *Proceedings of International Workshop on Acoustic Echo and Noise Control*.

Gradshteyn, I.S. and Ryzhik, I.M. (2000) *Table of Integrals Series and Products*, Academic Press, 6th edn..

Habets, E.A.P. (2007) *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*, Ph.D. thesis, Technische Universiteit Eindhoven.

Hendriks, R.C., Gerkmann, T., and Jensen, J. (2013) *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*, Morgan & Claypool.

Jensen, J. and Hendriks, R.C. (2012) Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions. *IEEE Transactions on Audio, Speech, and Language Processing*, **20** (1), 92 – 102.

Lebart, K., Boucher, J.M., and Denbigh, P.N. (2001) A new method based on spectral subtraction for speech dereverberation. *Acta Acustica*, **87**, 359–366.

Li, N. and Loizou, P.C. (2008) Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *Journal of the Acoustical Society of America*, **123** (3), 1673–1682.

Lotter, T. and Vary, P. (2005) Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model. *EURASIP Journal on Advances in Signal Processing*, **2005** (7), 1110–1126.

Madhu, N., Spriet, A., Jansen, S., Koning, R., and Wouters, J. (2013) The potential for speech intelligibility improvement using the ideal binary mask and the ideal Wiener filter in single channel noise reduction systems: Application to auditory prostheses. *IEEE Transactions on Audio, Speech, and Language Processing*, **21** (1), 61–70.

Martin, R. (2002) Speech enhancement using MMSE short time spectral estimation with

Gamma distributed speech priors, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 253–256.

Martin, R. (2005) Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Transactions on Speech and Audio Processing*, **13** (5), 845–856.

Martin, R. and Breithaupt, C. (2003) Speech enhancement in the DFT domain using Laplacian speech priors, in *Proceedings of International Workshop on Acoustic Echo and Noise Control*, pp. 87–90.

Martin, R. and Lotter, T. (2001) Optimal recursive smoothing of non-stationary periodograms, in *Proceedings of International Workshop on Acoustic Echo and Noise Control*, pp. 167–170.

Porter, J.E. and Boll, S.F. (1984) Optimal estimators for spectral restoration of noisy speech, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 18A.2.1–18A.2.4.

Roweis, S.T. (2001) One microphone source separation, in *Proceedings of Neural Information Processing Systems*, pp. 793–799.

Schreier, P.J. and Scharf, L.L. (2010) *Statistical Signal Processing of Complex-valued Data: The Theory of Improper and Noncircular Signals*, Cambridge University Press.

Vincent, E. (2007) Complex nonconvex l_p norm minimization for underdetermined source separation, in *Proceedings of International Conference on Independent Component Analysis and Signal Separation*, pp. 430–437.

Vincent, E. (2010) An experimental evaluation of Wiener filter smoothing techniques applied to under-determined audio source separation, in *Proceedings of International Conference on Latent Variable Analysis and Signal Separation*, pp. 157–164.

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

Wang, D.L. (2005) On ideal binary mask as the computational goal of auditory scene analysis, in *Speech Separation by Humans and Machines*, Springer, pp. 181–197.

Wolfe, P.J. and Godsill, S.J. (2003) Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. *EURASIP Journal on Applied Signal Processing*, **10**, 1043–1051.

You, C.H., Koh, S.N., and Rahardja, S. (2005) β-order MMSE spectral amplitude estimation for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, **13** (4), 475–486.