Chapter 3
Speech Dereverberation Using Statistical Reverberation Models*

Emanuël A.P. Habets

Abstract In speech communication systems, such as voice-controlled systems, hands-free mobile telephones and hearing aids, the received microphone signals are degraded by room reverberation, ambient noise and other interferences. This signal degradation can decrease the fidelity and intelligibility of speech and the word recognition rate of automatic speech recognition systems.

The reverberation process is often described using deterministic models that depend on a large number of unknown parameters. These parameters are often difficult to estimate blindly and are dependent on the exact spatial position of the source and receiver. In recently emerged speech dereverberation methods, which are feasible in practice, the reverberation process is described using a statistical model. This model depends on smaller number of parameters such as the reverberation time of the enclosure, which can be assumed to be independent of the spatial location of the source and receiver. This model can be utilized to estimate the spectral variance of part of the reverberant signal component. Together with an estimate of the spectral variance of the ambient noise, this estimate can then be used to enhance the observed noisy and reverberant speech.

In this chapter we provide a brief overview of dereverberation methods. We then describe single and multiple microphone algorithms that are able to jointly suppress reverberation and ambient noise. Finally, experimental results demonstrate the beneficial use of the algorithms developed.

Imperial College London, UK

* This research was supported by the Israel Science Foundation (grant no. 1085/05) and by the Technology Foundation STW, Applied Science Division of NWO and the Technology Programme of the Dutch Ministry of Economic Affairs.

3.1 Introduction

Speech signals that are received by a microphone at a distance from the speech source usually contain reverberation, ambient noise and other interferences. Reverberation is the process of multi-path propagation of an acoustic sound from its source to a microphone. The received microphone signal generally consists of a direct sound, reflections that arrive shortly after the direct sound (commonly called *early reverberation*) and reflections that arrive after the early reverberation (commonly called *late reverberation*). The combination of the direct sound and early reverberation is sometimes referred to as the *early speech component*. Early reverberation mainly contributes to spectral colouration, while late reverberation changes the waveform's temporal envelope as exponentially decaying tails are added at sound offsets. The colouration can be characterized by the spectral deviation σ, which is defined as the standard deviation of the log-amplitude frequency response of the Acoustic Impulse Response (AIR) [46].

For the development of dereverberation algorithms it is of great importance to have a good understanding of the effects of reverberation on speech perception. The reduction in speech intelligibly caused by late reverberation is especially noticeable for non-native listeners [72] and for listeners with hearing impairments [58]. The detrimental effects of reverberation on speech intelligibility have been attributed to two types of masking. Bolt and MacDonald [10] and Nábělek *et al.* [57] found evidence of *overlap-masking*, whereby late reverberation of a preceding phoneme masks a subsequent phoneme, and of *self-masking*, which refers to the time and frequency alterations of an individual phoneme.

In a reverberant room, speech intelligibility initially decreases with increasing source-microphone distance, but beyond the so-called critical distance speech intelligibility is approximately constant. The critical distance is the distance at which the direct-path energy is equal to the energy of all reflections. For an omnidirectional microphone the critical distance D_c is approximately given by [69]

$$D_{\mathrm{c}} = \sqrt{\frac{\ln(10^6)\,V}{4\pi\, c\, T_{60}}}, \tag{3.1}$$

where c is the sound velocity in ms^{-1}, V is the volume of the room in m^3 and T_{60} is the reverberation time in seconds. To obtain sufficiently intelligible speech it is typically recommended that the source-microphone distance is smaller than 0.3 times the critical distance. In a living room with dimensions 7 m $\times$ 5 m $\times$ 3 m and $T_{60} = 0.5$ s, the critical distance $D_c \approx 0.82$ m. Hence, the speech intelligibility would be affected even when the source-microphone distance is larger than 0.25 m.

Consonants play a more significant role in speech intelligibility than vowels. If the consonants are heard clearly, the speech can be understood more easily. In 1971 Peutz [60] proposed a measure called the articulation loss of consonants ($\mathrm{Al}_{\mathrm{cons}}$) that quantifies the reduction in perception of consonants due to reverberation. For distances smaller than the critical distance the measure depends on the source-

microphone distance, the reverberation time, and the volume of the room. Beyond the critical distance the measure depends only on the reverberation time. The speech intelligibility can be increased by decreasing the articulation loss, which can be achieved by decreasing the source-microphone distance or the reverberation time, or by increasing the room volume.

In 1982 Allen [4] reported a formula to predict the *subjective preference* of reverberant speech. The main result is given by the equation

$$P = P_{\max} - \sigma T_{60}, \tag{3.2}$$

where P is the subjective preference in some arbitrary units, $P_{\max}$ is the maximum possible preference, and σ is the spectral deviation in decibels (dB). According to this formula, decreasing either the spectral deviation σ or the reverberation time T_{60} results in an increased subjective preference of reverberant speech.

It would be convenient to assume that reverberation solely reduces intelligibility, but this assumption is incorrect [71]. Strong reflections that arrive shortly after the direct sound actually reinforce the direct sound and are therefore considered useful with regard to speech intelligibility. This reinforcement, which is often referred to as the *precedence effect*, is what makes it easier to hold conversations in closed rooms rather than outdoors.

While investigating the detrimental effects of reverberation on speech, it has become clear that the speech fidelity and intelligibility are mostly degraded by late reverberation. In addition, speech intelligibility is degraded by ambient noise. Therefore, we define the effective noise as the sum of the late reverberant component and the ambient noise component. In this chapter we describe a spectral enhancement method to suppress late reverberation and ambient noise, i.e., to estimate the early speech component. Due to the joint suppression of late reverberation and ambient noise, the effective noise is reduced and the fidelity and intelligibility of speech can be improved.

This chapter is organized as follows. In Sect. 3.2 a short review of dereverberation methods is provided. In Sect. 3.3 two statistical reverberation models are discussed. In Sect. 3.4 we derive a spectral estimator which can be used to jointly suppress late reverberation and ambient noise. In Sect. 3.5 we investigate the possibility of using multiple microphones in conjunction with spectral enhancement techniques for dereverberation. The spectral estimator derived in Sect. 3.4 requires an estimate of the spectral variance of the late reverberant signal component. In Sect. 3.6 such an estimator is derived using a statistical reverberation model. Estimation of the model parameters is discussed in Sect. 3.7. Experimental results that demonstrate the beneficial use of the described dereverberation methods are presented in Sect. 3.8. Finally, a summary and directions for further research are provided in Sect. 3.9.

3.2 Review of Dereverberation Methods

Reverberation reduction processes can be divided into many categories. They may, for example, be divided into single or multi-microphone techniques and into those primarily affecting colouration or those affecting late reverberation. We categorized the reverberation reduction processes depending on whether or not the AIR needs to be estimated. We then obtain two main categories, *viz. reverberation cancellation* and *reverberation suppression*.

3.2.1 Reverberation Cancellation

The first category, i.e., reverberation cancellation, consists of methods known as blind deconvolution. Much research has been undertaken on the topic of blind deconvolution; see [43] and the references therein. Multichannel methods appear particularly interesting because theoretically exact inverse-filtering can be achieved if the AIRs can be estimated and they do not have any common-zeros in the z-plane [56]. To achieve dereverberation without *a priori* knowledge of the room acoustics, many traditional methods assume that the source signal is independent and identically-distributed (i.i.d.). However, the i.i.d. assumption does not hold for speech-like signals. When applying such traditional deconvolution methods to speech, the speech generating process is deconvolved and the resulting speech signal is excessively whitened. Delcroix *et al.* proposed a method that consists of a multichannel equalizer and a compensation filter that reconstructs the colouration of the speech signal that is whitened by the equalizer [21]. Although perfect dereverberation is possible in theory, the method is sensitive to estimation errors of the covariance matrix that is required to compute the equalizer and the compensation filter. Another interesting method was developed by Gürelli and Nikias [33] and explores the null-space of the spatial correlation matrix, calculated from the received signals. It was shown that the null-space of the correlation matrix contains information on the acoustic transfer functions. This method has also potential in the speech processing framework and was extended by Gannot and Moonen [28]. In [44] the speech signal is modelled using a block stationary auto-regressive process while the room acoustics are modelled using an all-pole model. Bayesian parameter estimation techniques were then used to estimate the unknown parameters.

While good results can be achieved the methods in this category suffer from several limitations: (1) they have been shown to be insufficiently robust to small changes in the AIR [63, 73], (2) channels cannot be identified uniquely when they contain common zeros, (3) observation noise causes severe problems, and (4) some methods require knowledge of the order of the unknown system [45]. Detailed treatments on the problems involved are presented in Chaps. 5–7 and 9.

3.2.2 Reverberation Suppression

Methods in the second category, i.e., reverberation suppression, do not require an estimate of the AIR and explicitly exploit the characteristics of speech, the effect of reverberation on speech, or the characteristics of the AIR. Methods based on processing of the Linear Prediction (LP) residual signal belong to this category [30, 32, 78]. The peaks in the LP residual signal correspond to excitation events in voiced speech together with additional random peaks due to reverberation. These random peaks can be suppressed by, for example, averaging adjacent larynx-cycles, as proposed in [30].

Other, so-called, spatial processing methods use multiple microphones placed at different locations. They often use a limited amount of *a priori* knowledge of the AIR such as, for example, the direction of arrival of the desired source. The microphone signals can be processed to enhance or attenuate signals emanating from particular directions. The well-known delay and sum beamformer is a good example of such a method and belongs to the reverberation suppression category.

Recently, spectral enhancement methods have been used for speech dereverberation [37, 39, 41, 42, 49, 74]. Spectral enhancement of noisy speech has been a challenging problem for many researchers for over 30 years and is still an active research area, see, for example, [6, 17, 23, 24] and references therein. Spectral enhancement of noisy speech is often formulated as estimation of speech spectral components from a speech signal degraded by statistically independent additive noise. One of the earlier methods, and perhaps the most well-known method, is spectral subtraction [9, 50]. This method generally results in random narrow-band fluctuations in the residual noise, also known as musical tones, which are annoying and disturbing to the perception of the enhanced signal. Many variations have been developed to cope with musical tones [8, 9, 31, 36, 70]. Spectral subtraction makes minimal assumptions about the signal and noise, and when carefully implemented, it produces enhanced signals that may be acceptable for certain applications. Lebart *et al.* proposed a single-channel speech dereverberation method based on spectral subtraction to reduce the effect of overlap-masking [49]. The method estimates the short-term Power Spectral Density (PSD) of late reverberation based on a statistical reverberation model. This model exploits the fact that the envelope of the AIR decays exponentially and depends on a single parameter that is related to the reverberation time of the room. In [38] the authors showed that the estimated short-term PSD of late reverberation can be improved using multiple microphones. Additionally, the fine-structure of the speech signal is partly restored due to spatial averaging of the received power spectra.

A more advanced spectral enhancement method is the so-called statistical method, which is often designed to minimize the expected value of some distortion measure between the clean and estimated signals [11, 17, 25, 55]. This method requires reliable statistical models for the speech and noise signals, a perceptually meaningful distortion measure and an efficient signal estimator. A statistical speech model and perceptually meaningful distortion measure that are fully appropriate for spectral enhancement have not yet been determined. Hence, the variety of statistical

methods for spectral enhancement differ mainly in the statistical model [15, 25, 55], distortion measure [26, 52, 77] and the particular implementation of the spectral enhancement algorithm [23]. In this chapter we describe a statistical method for the enhancement of noisy and reverberant speech based on a Gaussian model for the speech and interferences and a squared error distortion measure.

3.3 Statistical Reverberation Models

Since the acoustic behaviour in real rooms is too complex to model explicitly, we make use of Statistical Room Acoustics (SRA). SRA provides a statistical description of the transfer function of the system between the source and the microphone in terms of a few key quantities, e.g., source-microphone distance, room volume and reverberation time. The crucial assumption of SRA is that the distribution of amplitudes and phases of individual plane waves, which sum up to produce sound pressure at some point in a room, is so close to random that the sound field is fairly uniformly distributed throughout the room volume. The validity of this description is subjected to a set of conditions that must be satisfied to ensure the accuracy of calculations. Our analysis therefore implicitly assumes that the following conditions hold [48, 63, 73]:

1. The dimensions of the room are relatively large compared to the longest wavelength of the sound of interest.
2. The average spacing of the resonance frequencies of the room must be smaller than one third of their bandwidth. In a room with volume V this condition is fulfilled for frequencies that exceed the Schroeder frequency: $f_{\mathrm{g}} = 2000\sqrt{T_{60}/V}$.
3. The source and the microphone are located in the interior of the room, at least a half-wavelength away from the walls.

3.3.1 Polack's Statistical Model

Sabine's [65] major contribution was the introduction of statistical methods to calculate the reverberation time of an enclosed space without considering the details of the space geometry. Schroeder extended Sabine's fundamental work [66, 67] and derived a frequency domain model and a set of statistical properties about the frequency response of the random impulse response.

Polack [61] developed a time-domain model complementing Schroeder's frequency domain model. In this model, an AIR is described as a realization of a non-stationary stochastic process. This model is defined as

$$h(n) = \begin{cases} b(n)\mathrm{e}^{-\bar{\zeta}n}, & \text{for } n \geq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{3.3}$$

where n denotes the discrete time index, $b(n)$ is a zero-mean stationary Gaussian noise sequence and $\bar{\zeta}$ is linked to the reverberation time T_{60} through

$$\bar{\zeta} \triangleq \frac{3\ln(10)}{T_{60} f_{\mathrm{s}}}, \tag{3.4}$$

where f_{s} denotes the sampling frequency in Hz. In contrast to the model in (3.3), the reverberation time is frequency dependent due to frequency dependent reflection coefficients of walls and other objects and the frequency dependent absorption coefficient of air [48]. This dependency can be taken into account by using a different model for each frequency band. In addition, it should be noted that Polack's statistical reverberation model is only valid in cases for which the distance between the source and the measurement point is greater than the critical distance D_c.

In the early 90s, Polack [62] proved that the most interesting properties of room acoustics are statistical when the number of 'simultaneously' arriving reflections exceeds a limit of about 10. In this case, the echo density is high enough such that the space can be considered to be in a fully diffused or mixed state. The essential requirement is ergodicity, which requires that any given reflection trajectory in the space will eventually reach all points. The ergodicity assumption is determined by the shape of the enclosure and the surface reflection properties. It should be noted that non-ergodic shapes will exhibit much longer mixing times and may not even have an exponential decay. Nevertheless, while it may not be true that all acoustic environments can be modelled using this statistical model, it is sufficiently accurate for most spaces.

The energy envelope of the AIR can be expressed as

$$\mathscr{E}\{h^2(n)\} = \sigma^2 \mathrm{e}^{-2\bar{\zeta}n}, \tag{3.5}$$

where σ^2 denotes the variance of $b(n)$, and $\mathscr{E}\{\cdot\}$ denotes spatial expectation. Here the spatial expectation is defined as the ensemble average over different realizations of the stochastic process in (3.3). Under the assumption that the space is ergodic, we may evaluate the ensemble average in (3.5) by spatial averaging so that different realizations of this stochastic process are obtained by varying either the position of the receiver or the source [47]. Note that the same stochastic process will be observed for all allowable positions (in terms of the third SRA condition) provided that the time origin is defined with respect to the signal emitted by the source and not with respect to the arrival time of the direct sound at the receiver.

3.3.2 Generalized Statistical Model

In many cases the source-microphone distance is smaller than the critical distance D_c, i.e., the Direct to Reverberant Ratio (DRR) is larger than 0 dB. In these cases Polack's statistical model, although useful when the source-microphone distance is larger than the critical distance, is not an accurate model of the AIR. In [39], a

generalized statistical model was proposed, which can be used when the source-microphone distance is smaller than the critical distance. To model the contribution of the direct-path, the AIR $h(n)$ is divided into two segments, *viz.* $h_{\mathrm{d}}(n)$ and $h_{\mathrm{r}}(n)$:

$$h(n) = \begin{cases} h_{\mathrm{d}}(n), & \text{for } 0 \leq n < n_{\mathrm{d}}, \\ h_{\mathrm{r}}(n), & \text{for } n \geq n_{\mathrm{d}}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.6}$$

The value n_{d} is chosen such that $h_{\mathrm{d}}(n)$ contains the direct-path and $h_{\mathrm{r}}(n)$ contains all reflections. Later we define the parameter n_{d} according to the frame rate of the time-frequency transformation. In practice, the direct-path is deterministic and could be modelled using a Dirac pulse. Unfortunately this would preclude us from creating a statistical model. To be able to model the energy related to the direct-path the following model is proposed:

$$h_{\mathrm{d}}(n) = b_{\mathrm{d}}(n)\mathrm{e}^{-\bar{\zeta}n}, \tag{3.7}$$

where $b_{\mathrm{d}}(n)$ is a white zero-mean Gaussian stationary noise sequence and $\bar{\zeta}$ is linked to the reverberation time T_{60} through (3.4). The reverberant component $h_{\mathrm{r}}(n)$ is described using the following model:

$$h_{\mathrm{r}}(n) = b_{\mathrm{r}}(n)\mathrm{e}^{-\bar{\zeta}n}, \tag{3.8}$$

where $b_{\mathrm{r}}(n)$ is a white zero-mean Gaussian stationary noise sequence. Under the SRA conditions the direct and reverberant component of the AIR are uncorrelated [63]. Therefore, it is reasonable to assume that $b_{\mathrm{d}}(n)$ and $b_{\mathrm{r}}(n)$ are uncorrelated, i.e., $\mathscr{E}\{b_{\mathrm{d}}(n)b_{\mathrm{r}}(n+\tau)\} = 0$ for $\tau \in \mathbb{Z}$.

The energy envelope of $h(n)$ can be expressed as

$$\mathscr{E}\{h^2(n)\} = \begin{cases} \sigma_{\mathrm{d}}^2\mathrm{e}^{-2\bar{\zeta}n}, & \text{for } 0 \leq n < n_{\mathrm{d}} \\ \sigma_{\mathrm{r}}^2\mathrm{e}^{-2\bar{\zeta}n}, & \text{for } n \geq n_{\mathrm{d}} \\ 0, & \text{otherwise,} \end{cases} \tag{3.9}$$

where σ_{d}^2 and σ_{r}^2 denote the variances of $b_{\mathrm{d}}(n)$ and $b_{\mathrm{r}}(n)$, respectively. When $\sigma_{\mathrm{d}}^2 < \sigma_{\mathrm{r}}^2$, the contribution of the direct-path can be neglected. Therefore, it is assumed that $\sigma_{\mathrm{d}}^2 \geq \sigma_{\mathrm{r}}^2$. Note that the generalized statistical model is equivalent to Polack's statistical model in the case $\sigma_{\mathrm{d}}^2 = \sigma_{\mathrm{r}}^2$.

3.4 Single-microphone Spectral Enhancement

In this section the spectral enhancement of a noisy and reverberant microphone signal is discussed. We start by formulating the spectral enhancement problem in Sect. 3.4.1. In Sect. 3.4.2 we show how the spectrum of the early speech component

can be estimated using the Minimum Mean Square Error (MMSE) Log Spectral Amplitude (LSA) estimator proposed by Cohen in [13]. This estimator depends on the so-called *a priori* Signal to Interference Ratio (SIR) that needs to be estimated in practice. In Sect. 3.4.3 we describe how the *a priori* SIR can be estimated.

3.4.1 Problem Formulation

The reverberant signal results from the convolution of the anechoic speech signal and a causal AIR. In this section we assume that the AIR is time-invariant and that its length is infinite. The reverberant speech signal at discrete-time n can be written as

$$z(n) = \sum_{l=-\infty}^{n} s(l)h(n-l). \tag{3.10}$$

To simplify the following discussion it is assumed that the direct sound arrives at time instance n, i.e., the direct-path is modelled by $h(0)$. It should be noted that this assumption can be made without loss of generality. Since our main goal is to suppress late reverberation we split the AIR into two components (see Fig. 3.1) such that

$$h(n) = \begin{cases} 0, & n < 0, \\ h_{\mathrm{e}}(n), & 0 \leq n < n_{\mathrm{e}}, \\ h_{\ell}(n), & n \geq n_{\mathrm{e}}, \end{cases} \tag{3.11}$$

where n_{e} is chosen such that $h_{\mathrm{e}}(n)$ consists of the direct-path and a few early reflections and $h_{\ell}(n)$ consists of all later reflections. The fraction $n_{\mathrm{e}}/f_{\mathrm{s}}$ can be used to define the time instance (relative to the time of arrival of the direct sound) from where the late reverberation is suppressed. Its value can be determined by the listener depending on his or her subjective preference but should be larger than the mixing time of the room, which is defined as the time it takes for initially adjacent sound rays to spread uniformly across the room [61]. In practice, $n_{\mathrm{e}}/f_{\mathrm{s}}$ usually ranges from 30 to 60 ms.

Using (3.11) we can write the microphone signal $x(n)$ as

$$x(n) = \underbrace{\sum_{l=n-n_{\mathrm{e}}+1}^{n} s(l)h_{\mathrm{e}}(n-l)}_{z_{\mathrm{e}}(n)} + \underbrace{\sum_{l=-\infty}^{n-n_{\mathrm{e}}} s(l)h_{\ell}(n-l)}_{z_{\ell}(n)} + v(n), \tag{3.12}$$

where $z_{\mathrm{e}}(n)$ is the early speech component, $z_{\ell}(n)$ denotes the late reverberant speech component, and $v(n)$ denotes the additive ambient noise component. The joint suppression of $z_{\ell}(n)$ and $v(n)$ decreases the effective noise level, and can increase the speech fidelity and intelligibility. Since the response of the first part of the AIR, i.e., $z_{\mathrm{e}}(n)$, remains unaltered we do not reduce the colourations caused by the early reflections.

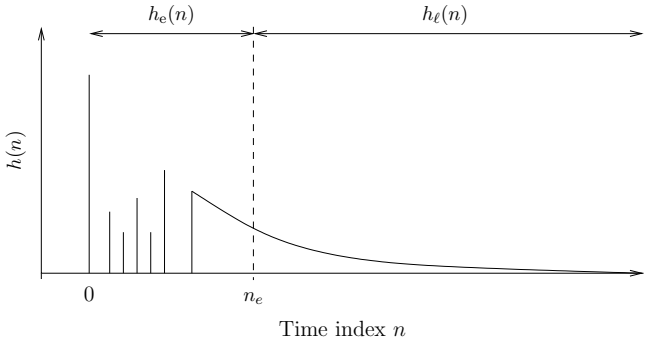


Fig. 3.1 Schematic representation of the acoustic impulse response

Estimating $z_e(n)$ is a challenging problem because both $s(n)$ and $h(n)$ are unknown. Here we formulate the problem of estimating $z_e(n)$, or in other words suppressing $z_\ell(n)$, using spectral enhancement. A block diagram of the spectral enhancement system is depicted in Fig. 3.2. The noisy and reverberant speech signal is denoted by $x(n)$, and is first transformed to the time-frequency domain by applying the short-time Fourier transform (STFT). Specifically,

$$X(\ell,k) = \sum_{n=0}^{K-1} x(n+\ell R)w(n)\mathrm{e}^{-j\frac{2\pi}{K}nk}, \tag{3.13}$$

where $j=\sqrt{-1}$, $w(n)$ is the analysis window of size K, and R is the number of samples separating two successive frames. The spectral component $X(\ell,k)$ can be used to estimate the spectral variance $\lambda_v(\ell,k) = E\{|V(\ell,k)|^2\}$ of the ambient noise and to estimate the spectral variance $\lambda_{z_\ell}(\ell,k) = E\{|Z_\ell(\ell,k)|^2\}$ of the late reverberant signal component $z_\ell(n)$. In the following, we assume that the spectral variance of the ambient noise is slowly time varying. Therefore, the spectral variance $\lambda_v(\ell,k)$ of the ambient noise can be estimated using the algorithm proposed by Martin in [54] or using the Improved Minima Controlled Recursive Averaging (IMCRA) algorithm proposed by Cohen in [14]. In contrast to $\lambda_v(\ell,k)$, the spectral variance $\lambda_{z_\ell}(\ell,k)$ of late reverberant signal component is highly time-varying due to the non-stationarity of the anechoic speech signal. In the application that is considered in this chapter, it is possible to estimate $\lambda_{z_\ell}(\ell,k)$ from the microphone signal. An estimator for $\lambda_{z_\ell}(\ell,k)$ is derived in Sect 3.6. For now, we assume that an estimate of the late reverberant spectral variance is available.

Using statistical signal processing, the spectral enhancement problem can be formulated as deriving an estimator $\hat{Z}_e(\ell,k)$ for the speech spectral coefficients such that the expected value of a certain distortion measure is minimized [17]. Let $\mathcal{H}_1(\ell,k)$ and $\mathcal{H}_0(\ell,k)$ denote the hypotheses for speech presence and absence in the spectral coefficient $Z_e(\ell,k)$, respectively. Such that

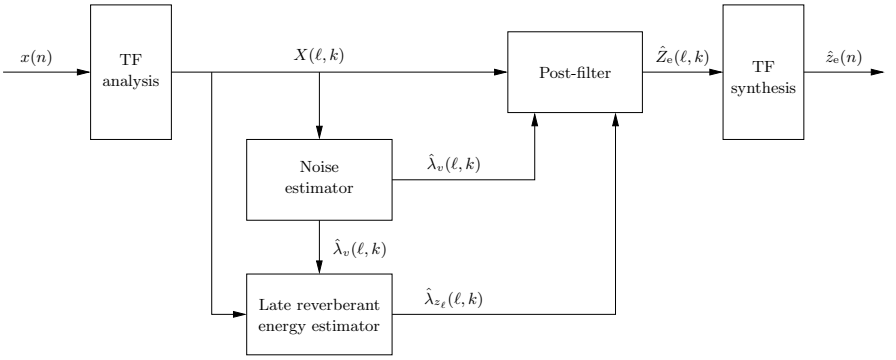


Fig. 3.2 Block diagram of the single-microphone spectral enhancement system for late reverberation and noise suppression

$$\mathcal{H}_1(\ell,k):\ X(\ell,k) = Z_{\mathrm{e}}(\ell,k) + Z_\ell(\ell,k) + V(\ell,k), \tag{3.14}$$

$$\mathcal{H}_0(\ell,k):\ X(\ell,k) = Z_\ell(\ell,k) + V(\ell,k). \tag{3.15}$$

Let $\hat{p}(\ell,k) = P(\mathcal{H}_1(\ell,k))$ denote an estimate for the probability that the desired speech component is present and let $\hat{\lambda}_{z_{\mathrm{e}}}(\ell,k)$ denote an estimate of the variance of the early speech spectral coefficient $Z_{\mathrm{e}}(\ell,k)$ under $\mathcal{H}_1(\ell,k)$. We can now calculate an estimator for $Z_{\mathrm{e}}(\ell,k)$ that minimizes the expected value of the distortion measure given $\hat{p}(\ell,k)$, $\hat{\lambda}_{z_{\mathrm{e}}}(\ell,k)$, the estimated late reverberant spectral variance $\hat{\lambda}_{z_\ell}(\ell,k) = E\{|\hat{Z}_\ell(\ell,k)|^2\}$, the estimated ambient noise spectral variance $\hat{\lambda}_v(\ell,k) = E\{|\hat{V}(\ell,k)|^2\}$ and the spectral coefficient $X(\ell,k)$:

$$\hat{Z}_{\mathrm{e}}(\ell,k) = \underset{\hat{Z}_{\mathrm{e}}(\ell,k)}{\operatorname{argmin}}\, E\left\{d\left(Z_{\mathrm{e}}(\ell,k), \hat{Z}_{\mathrm{e}}(\ell,k)\right)\right\}. \tag{3.16}$$

In the sequel we restrict ourselves to the squared error distortion measure, i.e.,

$$d\left(Z_{\mathrm{e}}(\ell,k), \hat{Z}_{\mathrm{e}}(\ell,k)\right) = \left|g(\hat{Z}_{\mathrm{e}}(\ell,k)) - \tilde{g}\left(Z_{\mathrm{e}}(\ell,k)\right)\right|^2, \tag{3.17}$$

where $g(Z_{\mathrm{e}})$ and $\tilde{g}(Z_{\mathrm{e}})$ are specific functions of Z_{e} that determine the fidelity criterion of the estimator. For the squared error distortion measure, the estimator $\hat{Z}_{\mathrm{e}}(\ell,k)$ is calculated from

$$\begin{aligned} g(\hat{Z}_{\mathrm{e}}(\ell,k)) &= E\left\{\tilde{g}(Z_{\mathrm{e}}(\ell,k)) \,\middle|\, X(\ell,k), \hat{p}(\ell,k)\right\} \\ &= \hat{p}(\ell,k)\, E\left\{\tilde{g}(Z_{\mathrm{e}}(\ell,k)) \,\middle|\, X(\ell,k), \mathcal{H}_1(\ell,k)\right\} \\ &\quad + (1-\hat{p}(\ell,k))\, E\left\{\tilde{g}(Z_{\mathrm{e}}(\ell,k)) \,\middle|\, X(\ell,k), \mathcal{H}_0(\ell,k)\right\}. \end{aligned} \tag{3.18}$$

Finally, given the estimated spectral component $\hat{Z}_{\mathrm{e}}(\ell,k)$ the early speech component $\hat{z}_{\mathrm{e}}(n)$ can be obtained using the inverse STFT,

$$\hat{z}_{\mathrm{e}}(n) = \sum_{\ell} \sum_{k=0}^{K-1} \hat{Z}_{\mathrm{e}}(\ell,k)\tilde{w}(n-\ell R)\mathrm{e}^{j\frac{2\pi}{K}k(n-\ell R)}, \tag{3.19}$$

where $\tilde{w}(n)$ is a synthesis window that satisfies the so-called completeness condition:

$$\sum_{\ell} w(n-\ell R)\tilde{w}(n-\ell R) = \frac{1}{K} \quad \text{for all } n. \tag{3.20}$$

Given analysis and synthesis windows that satisfy (3.20) we can reconstruct $\hat{z}(n)$ from its STFT coefficients $\hat{Z}(\ell,k)$. In practice, a Hamming window is often used for the synthesis window. A reasonable choice for the analysis window is the one with minimum energy [76], given by

$$w(n) = \frac{\tilde{w}(n)}{K\sum_{\ell}\tilde{w}^2(n-\ell R)}. \tag{3.21}$$

The inverse STFT is efficiently implemented using the weighted overlap-add method [20].

3.4.2 MMSE Log-spectral Amplitude Estimator

In the previous Section it was shown that the received microphone signal is degraded by late reverberation and ambient noise. In this section, a spectral amplitude estimator is developed that can be used to estimate the early spectral speech component $Z_{\mathrm{e}}(\ell,k)$ in the presence of late reverberation and ambient noise.

While there are many fidelity criteria that are of interest for speech enhancement it has been found that the MMSE of the log-spectral amplitude is advantageous to other MMSE estimators in the case of noise suppression [17]. The MMSE-LSA estimator is found by using the following functions:

$$g(\hat{Z}_{\mathrm{e}}(\ell,k)) = \log_{\mathrm{e}}(|\hat{Z}_{\mathrm{e}}(\ell,k)|), \tag{3.22}$$

$$\tilde{g}(Z_{\mathrm{e}}(\ell,k)) = \begin{cases} \log_{\mathrm{e}}(|Z_{\mathrm{e}}(\ell,k)|) & \text{under } \mathcal{H}_1(\ell,k) \\ \log_{\mathrm{e}}(G_{\min}(\ell,k)\,|X(\ell,k)|) & \text{under } \mathcal{H}_0(\ell,k). \end{cases} \tag{3.23}$$

The MMSE-LSA estimator is obtained by substituting (3.22) and (3.23) into (3.18). Using a Gaussian model for the spectral coefficients, the MMSE-LSA gain function yields [13]

$$G_{\text{MMSE-LSA}}(\ell,k) = \{G_{\text{LSA}}(\ell,k)\}^{p(\ell,k)}\,\{G_{\min}(\ell,k)\}^{1-p(\ell,k)}, \tag{3.24}$$

where $G_{\text{LSA}}(\ell,k)$ is the LSA gain function derived by Ephraim and Malah [26] and $G_{\min}(\ell,k)$ is the lower bound for the gain when the signal is absent and specifies the maximum amount of suppression in those frames. An efficient estimator for the speech presence probability $\hat{p}(\ell,k)$ was developed in [13]. Let $\xi(\ell,k)$ denote the *a*

priori SIR,

$$\xi(\ell,k) = \frac{\lambda_{z_\mathrm{e}}(\ell,k)}{\lambda_{z_\ell}(\ell,k) + \lambda_v(\ell,k)}, \tag{3.25}$$

and $\gamma(\ell,k)$ denote the *a posteriori* SIR,

$$\gamma(\ell,k) = \frac{|X(\ell,k)|^2}{\lambda_{z_\ell}(\ell,k) + \lambda_v(\ell,k)}. \tag{3.26}$$

Here $X(\ell,k)$ denotes the spectral coefficient of the microphone signal and $\lambda_{z_\mathrm{e}}(\ell,k)$, $\lambda_{z_\ell}(\ell,k)$, and $\lambda_v(\ell,k)$ denote the spectral variances of the early speech component, late reverberation, and ambient noise, respectively. While the *a posteriori* SIR can be calculated directly, the *a priori* SIR cannot because the spectral variance $\lambda_{z_\mathrm{e}}(\ell,k)$ of the early speech component in (3.25) is unobservable. The estimation of the *a priori* SIR is treated in Section 3.4.3.

The LSA gain function depends on the *a posteriori* and *a priori* SIR and is given by [26]

$$G_{\mathrm{LSA}}(\ell,k) = \frac{\xi(\ell,k)}{1+\xi(\ell,k)} \exp\left(\frac{1}{2}\int_{\zeta(\ell,k)}^{\infty} \frac{e^{-t}}{t}\,\mathrm{d}t\right), \tag{3.27}$$

where

$$\zeta(\ell,k) = \frac{\xi(\ell,k)}{1+\xi(\ell,k)}\,\gamma(\ell,k). \tag{3.28}$$

To avoid speech distortions $G_{\min}$ is usually set between -12 and -18 dB. However, in practice the late reverberation plus ambient noise needs to be reduced more than 12–18 dB. Therefore, we like to control the maximum suppression of the late reverberant speech component and ambient noise separately. Due to the time-varying nature of the interferences the lower-bound becomes time and frequency dependent. Under the assumption that the interferences are uncorrelated a modified lower-bound is given by

$$G_{\min}(\ell,k) = \frac{G_{\min,z_\ell}\,\hat{\lambda}_{z_\ell}(\ell,k) + G_{\min,v}\,\hat{\lambda}_v(\ell,k)}{\hat{\lambda}_{z_\ell}(\ell,k) + \hat{\lambda}_v(\ell,k)}, \tag{3.29}$$

where $G_{\min,z_\ell}$ and $G_{\min,v}$ are used to control the maximum suppression of late reverberation and ambient noise, respectively. When $G_{\min,z_\ell} = 0$ the late reverberation is suppressed down to the residual level of the ambient noise, as shown in [40]. The results of an informal listening test using stationary ambient noise confirmed that the sound level of the residual interference was stationary in case the modified lower-bound $G_{\min}(\ell,k)$ was used, while the sound level of the residual interference fluctuated when the constant lower bound $G_{\min}$ was used.

An estimate of the early spectral speech component $Z_\mathrm{e}(\ell,k)$ can now be obtained using the amplitude estimate and the phase of the noisy and reverberant spectral coefficient $X(\ell,k)$, i.e.,

$$\hat{Z}_\mathrm{e}(\ell,k) = G_{\text{MMSE-LSA}}(\ell,k)\,X(\ell,k). \tag{3.30}$$

3.4.3 *a priori* SIR Estimator

In this section we focus on the *a priori* SIR estimation. The *a priori* SIR in (3.25) can be written as

$$\frac{1}{\xi(\ell,k)} = \frac{1}{\xi_{z_\ell}(\ell,k)} + \frac{1}{\xi_v(\ell,k)}, \tag{3.31}$$

with

$$\xi_\vartheta(\ell,k) = \frac{\lambda_{z_e}(\ell,k)}{\lambda_\vartheta(\ell,k)}, \tag{3.32}$$

where $\vartheta \in \{z_\ell, v\}$. Hence, the total *a priori* SIR can be calculated using the *a priori* SIRs of each interference separately [34, 35, 40]. By doing this, one gains control over (1) the trade-off between the interference reduction and the distortion of the desired signal, and (2) the *a priori* SIR estimation approach for each interference. In some cases, it might be desirable to reduce one of the two interferences at the cost of larger speech distortion, while reducing the other interference less to avoid distortion. In this Section it is shown how the decision-directed approach, proposed by Ephraim and Malah in [25], can be used to estimate the individual *a priori* SIRs.

In the case when the early speech component and the late reverberant signal are very small, the *a priori* SIRs $\xi_{z_\ell}(\ell,k)$ may be unreliable since $\lambda_{z_e}(\ell,k)$ and $\lambda_{z_\ell}(\ell,k)$ are close to zero. In the following, we assume that there is always a certain amount of ambient noise, i.e., $\lambda_v(\ell,k) > 0$. We propose to calculate $\xi(\ell,k)$ using only the most important and reliable *a priori* SIRs as follows:

$$\xi(\ell,k) = \begin{cases} \xi_v(\ell,k), & 10\log_{10}\left(\frac{\lambda_v(\ell,k)}{\lambda_{z_\ell}(\ell,k)}\right) > \beta^{\mathrm{dB}}, \\ \dfrac{\xi_{z_\ell}(\ell,k)\,\xi_v(\ell,k)}{\xi_{z_\ell}(\ell,k)+\xi_v(\ell,k)}, & \text{otherwise}, \end{cases} \tag{3.33}$$

where the threshold β^{dB} specifies the level difference between $\lambda_v(\ell,k)$ and $\lambda_{z_\ell}(\ell,k)$ in dB. When the noise power level is β^{dB} higher than the late reverberant power level, the total *a priori* SIR, $\xi(\ell,k)$, will be equal to $\xi_v(\ell,k)$. Otherwise $\xi(\ell,k)$ will depend on both $\xi_v(\ell,k)$ and $\xi_{z_\ell}(\ell,k)$.

The decision-directed based estimator [12, 25] is given by

$$\hat{\xi}(\ell,k) = \max\left\{\eta\,\frac{G^2_{\mathrm{LSA}}(\ell-1,k)\,|X(\ell-1,k)|^2}{\lambda_{z_\ell}(\ell,k)+\lambda_v(\ell,k)} + (1-\eta)\,\psi(\ell,k), \xi_{\min}\right\}, \tag{3.34}$$

where $\psi(\ell,k) = \gamma(\ell,k) - 1$ is the *instantaneous* SIR, $\gamma(\ell,k)$ is the *a posteriori* SIR as defined in (3.26), and $\xi_{\min}$ is a lower-bound on the *a priori* SIR that controls the residual interference level when hypothesis $\mathcal{H}_1$ is assumed to be true (i.e., when the desired speech is assumed to be active). The weighting factor η $(0 \le \eta \le 1)$ controls the tradeoff between the amount of noise reduction and distortion [12, 25]. To estimate $\xi_\vartheta(\ell,k)$ we use the following expression:

$$\hat{\xi}_{\vartheta}(\ell,k) = \max\left\{\eta_{\vartheta}\frac{G^2_{\mathrm{LSA}}(\ell-1,k)\,|X(\ell-1,k)|^2}{\lambda_{\vartheta}(\ell-1,k)} + (1-\eta_{\vartheta})\;\psi_{\vartheta}(\ell,k), \xi_{\min,\vartheta}\right\}, \tag{3.35}$$

where

$$\begin{aligned}\psi_{\vartheta}(\ell,k) &= \frac{\lambda_{z_\ell}(\ell,k)+\lambda_v(\ell,k)}{\lambda_{\vartheta}(\ell,k)}\psi(\ell,k)\\ &= \frac{|Y(\ell,k)|^2 - \left[\lambda_{z_\ell}(\ell,k)+\lambda_v(\ell,k)\right]}{\lambda_{\vartheta}(\ell,k)},\end{aligned} \tag{3.36}$$

and $\xi_{\min,\vartheta}$ is the lower bound on the *a priori* SIR $\xi_{\vartheta}(\ell,k)$.

3.5 Multi-microphone Spectral Enhancement

Single-microphone systems only exploit the spectral and temporal diversity of the received signal. Reverberation and most ambient noise sources, of course, also induce spatial diversity. To be able to additionally exploit this diversity, multiple microphones must be used and their outputs must be combined by a suitable spatial processor, e.g., a delay-and-sum beamformer, a filter-and-sum beamformer or an adaptive beamformer. Although spatial processors yield a significant improvement of the speech quality, the reverberation suppression is limited and the noise suppression is insufficient when the noise field is non-coherent or diffuse. In addition to the spatial processor a single-channel post-filter should be used to achieve satisfactory results.

In this section we will elaborate on the use of multiple microphone signals for speech dereverberation. In Sect. 3.5.1 we formulate the multi-microphone speech dereverberation problem. In Sect. 3.5.2 we describe two multi-microphone speech enhancement systems for ambient noise and reverberation suppression. In Sect. 3.5.3 we propose a method to enhance the speech presence probability estimation when multiple microphone signals are available.

3.5.1 Problem Formulation

In Sect. 3.4 we exploited the spectral and temporal diversity of the received signal to estimate the early speech component using a single microphone signal. When the signals of multiple microphones are combined using a suitable spatial processor it is possible to 'focus' on the desired source. The effect of early and late reflections can be suppressed to a degree depending on the spatial processor employed.

The reverberant signal at the m^{th} microphone results from the convolution of the anechoic speech signal $s(n)$ and a causal AIR $h_m(n)$. Here we assume that the AIR is time-invariant and that its length is infinite. The reverberant speech signal at

discrete-time n can be written as

$$z_m(n) = \sum_{l=0}^{\infty} h_m(l)\, s(n-l). \tag{3.37}$$

The m^{th} microphone signal is given by

$$x_m(n) = z_m(n) + v_m(n), \tag{3.38}$$

where $v_m(n)$ denote the additive ambient noise received by the m^{th} microphone.

In the STFT domain we can write (3.38) as

$$X_m(\ell,k) = Z_{\text{e},m}(\ell,k) + Z_{\ell,m}(\ell,k) + V_m(\ell,k), \tag{3.39}$$

where $Z_{\text{e},m}(\ell,k)$, $Z_{\ell,m}(\ell,k)$, and $V_m(\ell,k)$ denote the early and late spectral speech components and the ambient noise at the m^{th} microphone, respectively.

Our objective is to obtain an estimate of the early speech component without using detailed knowledge of the AIRs. Instead of estimating $Z_{\text{e},m}(\ell,k)$ with $m \in \{1,\ldots,M\}$, we propose to estimate a spatially filtered version of all early speech components.

3.5.2 Two Multi-microphone Systems

In this section we describe two multi-microphone systems that can be used to suppress ambient noise and reverberation. The first system consists of a Minimum Variance Distortionless Response (MVDR) beamformer followed by a single-channel post-filter. The second system consists of a non-linear spatial processor followed by a single-channel post-filter that was especially designed for speech dereverberation in [39].

3.5.2.1 MVDR Beamformer and Single-channel MMSE Estimator

This multi-microphone system consists of two stages. First, an MVDR beamformer is applied to the microphone signals. Second, a single-channel MMSE estimator is applied to the output of the MVDR beamformer.

Let us define $\mathbf{X}(\ell,k) = [X_1(\ell,k), X_2(\ell,k), \ldots, X_M(\ell,k)]^T$ and $\mathbf{V}(\ell,k) = [V_1(\ell,k), V_2(\ell,k), \ldots, V_M(\ell,k)]^T$. The MVDR filter, denoted by $\mathbf{W}(\ell,k) = [W_1(\ell,k), W_2(\ell,k), \ldots, W_M(\ell,k)]^T$, is found by solving the following minimization problem:

$$\mathbf{W}_{\mathrm{MVDR}}(\ell,k) = \underset{\mathbf{W}(k)}{\operatorname{argmin}} \left\{ (\mathbf{W}(k))^H \mathbf{\Lambda}_{\mathbf{VV}}(\ell,k)\mathbf{W}(k) \right\}$$
$$\text{subject to } (\mathbf{W}(k))^H \mathbf{C}(k) = 1, \quad (3.40)$$

where $(\cdot)^H$ denotes the Hermitian transpose, $\mathbf{\Lambda}_{\mathbf{VV}}(\ell,k) = E\{\mathbf{V}(\ell,k)\mathbf{V}^H(\ell,k)\}$ denotes the spatial PSD matrix of the noise, and $\mathbf{C}(k)$ denotes a pre-defined constraint column vector of length M.

A major question remains how to define the constraint $\mathbf{C}(k)$ and thereby the signal which is undistorted by the MVDR beamformer. One solution would be to estimate the reverberant speech component $Z_m(\ell,k)$ for $m \in \{1,\ldots,M\}$ (see, for example, [27]). In this case, the beamformer only reduces noise (and therefore no reverberation). Here we chose to align the direct sound signals of the desired source at the output of the MVDR beamformer. Due to the spatial directivity of the beamformer the spectral coloration induced by early reflections is slightly reduced.

Let us assume that the desired source is located in the far-field, such that the propagation of the direct sound can be modelled by $\mathbf{H}_{\mathrm{d}}(k) = \mathrm{e}^{-j\omega_k\tau_1}\tilde{\mathbf{H}}_{\mathrm{d}}(k)$, where $\tilde{\mathbf{H}}_{\mathrm{d}}(k) = [1, \mathrm{e}^{-j\omega_k\tau_{12}}, \ldots, \mathrm{e}^{-j\omega_k\tau_{1M}}]^T$, $\omega_k = 2\pi f_{\mathrm{s}}k/K$, τ_1 denotes the propagation time of the desired source signal to the first microphone and τ_{1m} $(2 \leq m \leq M)$ denotes the relative delay [also known as time difference of arrival (TDOA)] of the desired source signal between the m^{th} and the first microphone. The aim of the constraint of the MVDR beamformer is to align the direct-paths of the desired source at the output of the MVDR beamformer. Therefore, the constraint vector $\mathbf{C}(k)$ can be defined as

$$\mathbf{C}(k) = \tilde{\mathbf{H}}_{\mathrm{d}}(k). \quad (3.41)$$

Estimation of the TDOAs is beyond the scope of this chapter in which we assume that the TDOAs are known.

The solution of the minimization problem in (3.40) is given by

$$\mathbf{W}_{\mathrm{MVDR}}(\ell,k) = \frac{\mathbf{\Lambda}_{\mathbf{VV}}^{-1}(\ell,k)\mathbf{C}(k)}{\mathbf{C}^H(k)\mathbf{\Lambda}_{\mathbf{VV}}^{-1}(\ell,k)\mathbf{C}(k)}. \quad (3.42)$$

The output of the MVDR beamformer is given by

$$\begin{aligned} Q(\ell,k) &= (\mathbf{W}_{\mathrm{MVDR}}(\ell,k))^H \mathbf{X}(\ell,k) \\ &= Q_z(\ell,k) + Q_v(\ell,k), \end{aligned} \quad (3.43)$$

where $Q_z(\ell,k)$ and $Q_v(\ell,k)$ denote the residual reverberant and noise component at the beamformer's output. The spectral variance of $Q(\ell,k)$ is given by

$$\lambda_q(\ell,k) = E\{Q(\ell,k)(Q(\ell,k))^*\} \quad (3.44)$$
$$= \lambda_{q_z}(\ell,k) + \lambda_{q_v}(\ell,k), \quad (3.45)$$

where $(\cdot)^*$ denotes the complex conjugate, $\lambda_{q_z}(\ell,k)$ and $\lambda_{q_v}(\ell,k)$ denote the spectral variances of the residual reverberant and noise component at the beamformer's output. In addition, we can express $\lambda_{q_z}(\ell,k)$ as

$$\begin{aligned}\lambda_{q_z}(\ell,k) &= E\{Q_z(\ell,k)(Q_z(\ell,k))^*\} \\ &= \lambda_{q_e}(\ell,k) + \lambda_{q_\ell}(\ell,k),\end{aligned} \tag{3.46}$$

where $\lambda_{q_e}(\ell,k)$ and $\lambda_{q_\ell}(\ell,k)$ denote the residual early and late reverberation at the output of the beamformer. The spectral variance of the noise at the output of the MVDR beamformer is given by

$$\lambda_{q_v}(\ell,k) = \frac{1}{\mathbf{C}^H(k)\mathbf{\Lambda}_{\mathbf{VV}}^{-1}(\ell,k)\mathbf{C}(k)}. \tag{3.47}$$

Assuming that the residual early and late reverberant signal components are mutually uncorrelated we can reduce the residual late reverberation at the output of the MVDR beamformer using a spectral enhancement technique.

Let us now consider the case in which the ambient noise field is spatially white, i.e., $\mathbf{\Lambda}_{\mathbf{VV}}(\ell,k) = \sigma_v^2\,\mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix. In this case the MVDR beamformer reduces to the well-known delay and sum beamformer, i.e.,

$$\mathbf{W}_{\mathrm{MVDR}}(\ell,k) = \frac{1}{M}\tilde{\mathbf{H}}_d(\ell,k). \tag{3.48}$$

Although the output of the beamformer is not completely dereverberated the signal will contain less reverberation than any one of the observed microphone signals. Using statistical room acoustics, Gaubitch and Naylor derived an analytic expression to calculate the DRR improvement of the delay and sum beamformer compared to the best microphone [29]. Their result demonstrates that the reverberation reduction of the delay and sum beamformer is limited, especially when the source-microphone distance is larger than the critical distance.

Here we employ a single-channel MMSE log spectral amplitude estimator as described in Sect. 3.4 to estimate the residual early speech component at the beamfomer's output. In order to compute the LSA gain function (3.27) we redefine the *a priori* and *a posteriori* SIR as

$$\xi(\ell,k) = \frac{\lambda_{q_e}(\ell,k)}{\lambda_{q_\ell}(\ell,k) + \lambda_{q_v}(\ell,k)} \tag{3.49}$$

and

$$\gamma(\ell,k) = \frac{|Q(\ell,k)|^2}{\lambda_{q_\ell}(\ell,k) + \lambda_{q_v}(\ell,k)}, \tag{3.50}$$

respectively. The spectral variance $\lambda_{q_v}(\ell,k)$ of the residual noise can be estimated either by estimating $\mathbf{\Lambda}_{\mathbf{VV}}(\ell,k)$ during noise only periods and using (3.47) or by using a minimum statistics approach [14, 54]. The late reverberant spectral variance $\lambda_{q_\ell}(\ell,k)$ can be obtained from $Q(\ell,k)$ in a similar way to how $\lambda_{z_\ell}(\ell,k)$ can be obtained from $Z(\ell,k)$, as described in Sect. 3.6.

3.5.2.2 Non-linear Spatial Processor

In [39] it was shown that the output signal of the delay and sum beamformer may contain undesired signal components that result from the spatial correlation between the acoustic channels. The spatial correlation mainly causes problems at low frequencies and becomes more severe when the inter-microphone distance is small. To avoid the creation of these undesired components, a non-linear spatial processor was proposed that can be used when the noise field is spatially white. The spatial processor computes the amplitude and phase spectrum independently. Firstly, the observed spectra are delayed according to the DOA of the desired source. Secondly, the amplitude spectrum is computed from the squared value of the average PSDs:

$$Q(\ell,k) = \left(\frac{1}{M} \sum_{m=1}^{M} \left| X_m(\ell,k) \mathrm{e}^{j\omega_k \tau_{1m}} \right|^2 \right)^{\frac{1}{2}}, \tag{3.51}$$

where τ_{1m} denotes the TDOA of the desired source signal between the m^{th} and the first microphone (by definition $\tau_{11} = 0$). Finally, the phase spectrum is computed by averaging the phase spectra of the properly delayed signals:

$$\varphi(\ell,k) = \arg \left\{ \frac{1}{M} \sum_{m=1}^{M} X_m(\ell,k) \mathrm{e}^{j\omega_k \tau_{1m}} \right\}. \tag{3.52}$$

It should be noted that the phase spectrum is equal to the phase spectrum of the delay and sum beamformer. The output of the non-linear spatial processor is given by

$$Y_{\text{NL}}(\ell,k) = Q(\ell,k) \mathrm{e}^{j\varphi(\ell,k)}. \tag{3.53}$$

Due to the averaging of the PSDs the proposed spatial processor is unable to reduce any noise. The PSD of the noise in $Y_{\text{NL}}(\ell,k)$ is given by $\frac{1}{M} \sum_{m=1}^{M} |V_m(\ell,k)|^2$.

We can now apply the single-microphone spectral enhancement algorithm that was described in Sect. 3.4 to $Y_{\text{NL}}(\ell,k)$. The spectral variance $\lambda_{z_\ell}(\ell,k)$ of the late reverberant speech component can be estimated using $Y_{\text{NL}}(\ell,k)$ in a way similar to how $\lambda_{z_\ell}(\ell,k)$ can be estimated from $X(\ell,k)$. Using statistical room acoustics it can be shown that the expected value of the spatially averaged acoustic transfer functions is white. Since the statistical reverberation models in Sect. 3.3 are based on this assumption, the result obtained sounds better than the single-microphone spectral enhancement. Furthermore, due to the spatial averaging, the spectral colouration that is caused by the early reflections is slightly reduced.

3.5.3 Speech Presence Probability Estimator

In order to compute the MMSE-LSA gain function (3.24) we require an estimate of the *a posteriori* speech presence probability $p(\ell,k)$. The *a posteriori* speech pres-

ence probability $p(\ell,k)$ can be obtained from Bayes' rule, which, under a Gaussian model for the spectral coefficients, reduces to [13]

$$p(\ell,k) = \left\{1 + \frac{1-p(\ell,k|\ell-1)}{p(\ell,k|\ell-1)}\,(1+\xi(\ell,k))\exp\left(-\zeta(\ell,k)\right)\right\}^{-1}, \tag{3.54}$$

where $p(\ell,k|\ell-1)$ denotes the *a priori* speech presence probability, $\xi(\ell,k)$ is the *a priori* SIR and $\zeta(\ell,k)$ is defined in (3.28). In this section we develop an efficient estimator for the *a priori* speech presence probability $p(\ell,k|\ell-1)$, which exploits the strong correlation of speech presence in neighbouring frequency bins of consecutive frames and the strong spatial coherence of the desired signal.

We propose to estimate the *a posteriori* speech presence probability using four probabilities that are obtained using a soft-decision approach. Three probabilities, i.e., $P_{\text{local}}(\ell,k)$, $P_{\text{global}}(\ell,k)$, and $P_{\text{frame}}(\ell)$, are proposed by Cohen in [13], and are based on the time-frequency distribution of the estimated *a priori* SIR, $\xi(\ell,k)$. These probabilities reflect the strong correlation of speech presence in neighbouring frequency bins of consecutive frames. Since the spatial coherence of the desired direct sound is much larger than the spatial coherence of the reverberant sound, we propose to relate the fourth probability, denoted by $P_{\text{spatial}}(\ell,k)$, to the spatial coherence of the received signals. In [42] we proposed to determine $P_{\text{spatial}}(\ell,k)$ using Mean Square Coherence (MSC). Firstly, we smooth the MSC estimate in time and frequency to reduce its variance. Secondly, we map the MSC value to the probability $P_{\text{spatial}}(\ell,k)$. The latter can easily be achieved since the MSC value lies between zero and one.

The MSC is defined as

$$\Phi_{\text{MSC}}(\ell,k) \triangleq \frac{|\lambda_{x_{21}}(\ell,k)|^2}{\lambda_{x_1}(\ell,k)\lambda_{x_2}(\ell,k)}, \tag{3.55}$$

where $\lambda_{x_{21}}(\ell,k) = E\{X_2(\ell,k)(X_1(\ell,k))^*\}$ denotes the cross spectral density, and $\lambda_{x_1}(\ell,k)$ and $\lambda_{x_2}(\ell,k)$ are the power spectral densities. In addition, we know that $0 \leq \Phi_{\text{MSC}}(\ell,k) \leq 1$.

Let η $(0 \leq \eta_{\text{s}} \leq 1)$ denote a smoothing parameter. Then, the power and cross spectral density are estimated using

$$\hat{\lambda}_{x_i}(\ell,k) = \eta_{\text{s}}\hat{\lambda}_{x_i}(\ell-1,k) + (1-\eta_{\text{s}})|X_i(\ell,k)|^2, \quad i \in \{1,2\} \tag{3.56}$$

and

$$\hat{\lambda}_{x_{21}}(\ell,k) = \eta_{\text{s}}\hat{\lambda}_{x_{21}}(\ell-1,k) + (1-\eta_{\text{s}})X_2(\ell,k)(X_1(\ell,k))^*, \tag{3.57}$$

respectively. The MSC is further smoothed over different frequencies using

$$\tilde{\Phi}_{\text{MSC}}(\ell,k) = \sum_{i=-w_{\text{MSC}}}^{w_{\text{MSC}}} b(i)\Phi_{\text{MSC}}(\ell,k+i), \tag{3.58}$$

where $b(i)$ are the coefficients of a normalized window ($\sum_{i=-w_{\mathrm{MSC}}}^{w_{\mathrm{MSC}}} b(i) = 1$) of size $2w_{\mathrm{MSC}}+1$ that determine the frequency smoothing. In the case when more than two microphone signals are available one could average the MSC over different microphone pairs (with equal inter-microphone distance) to improve the estimation procedure even further.

The spatial speech presence probability $\hat{P}_{\mathrm{spatial}}(\ell,k)$ is related to (3.58) by

$$\hat{P}_{\mathrm{spatial}}(\ell,k) = \begin{cases} 0, & \text{for } \tilde{\Phi}_{\mathrm{MSC}}(\ell,k) \leq \Phi_{\min}, \\ 1, & \text{for } \tilde{\Phi}_{\mathrm{MSC}}(\ell,k) \geq \Phi_{\max}, \\ \frac{\tilde{\Phi}_{\mathrm{MSC}}(\ell,k)-\Phi_{\min}}{\Phi_{\max}-\Phi_{\min}}, & \text{otherwise}, \end{cases} \tag{3.59}$$

where $\Phi_{\min}$ and $\Phi_{\max}$ are the minimum and maximum threshold values for $\tilde{\Phi}_{\mathrm{MSC}}(\ell,k)$, respectively.

Finally, an estimate of the *a priori* speech presence probability is obtained by

$$\hat{p}(\ell,k|\ell-1) = \hat{P}_{\mathrm{local}}(\ell,k)\hat{P}_{\mathrm{global}}(\ell,k)\hat{P}_{\mathrm{frame}}(l)\hat{P}_{\mathrm{spatial}}(\ell,k). \tag{3.60}$$

3.6 Late Reverberant Spectral Variance Estimator

In this section we derive a spectral variance estimator for the late reverberant spectral component, $Z_\ell(\ell,k)$, using the generalized statistical reverberation model described in Sect. 3.3.

Before the spectral variance $\lambda_{z_\ell}(\ell,k) = E\{|Z_\ell(\ell,k)|^2\}$ can be estimated, we need to obtain an estimate of the spectral variance of the reverberant spectral component $Z(\ell,k)$ denoted by $\lambda_z(\ell,k)$. Assuming that the spectral coefficients of the reverberant signal and the noise are mutually independent Gaussian random variables, an estimate of the spectral variance $\lambda_z(\ell,k)$ can be obtained by calculating the following conditional expectation:

$$\begin{aligned} \hat{\lambda}_z(\ell,k) &= E\{|Z(\ell,k)|^2 \,|\, X(\ell,k)\} \\ &= |G_{\mathrm{SP}}(\ell,k)X(\ell,k)|^2, \end{aligned} \tag{3.61}$$

where $G_{\mathrm{SP}}(\ell,k)$ denotes the MMSE spectral power gain function. This gain function is given by [3]

$$G_{\mathrm{SP}}(\ell,k) = \frac{\xi_{\mathrm{SP}}(\ell,k)}{1+\xi_{\mathrm{SP}}(\ell,k)} \left(\frac{1}{\gamma_{\mathrm{SP}}(\ell,k)} + \frac{\xi_{\mathrm{SP}}(\ell,k)}{1+\xi_{\mathrm{SP}}(\ell,k)} \right), \tag{3.62}$$

where

$$\xi_{\mathrm{SP}}(\ell,k) = \frac{\lambda_z(\ell,k)}{\lambda_v(\ell,k)} \tag{3.63}$$

and

$$\gamma_{\mathrm{SP}}(\ell,k) = \frac{|X(\ell,k)|^2}{\lambda_v(\ell,k)} \tag{3.64}$$

denote the *a priori* and *a posteriori* SIRs, respectively. The *a priori* SIR is estimated using the decision-directed approach proposed by Ephraim and Malah [25]. Estimates of the spectral variance, $\lambda_v(\ell,k)$, of the noise in the received signal $x(n)$ can be estimated using so-called minimum statistics approaches [14, 54].

In order to derive an estimator for the spectral variance of the late reverberant signal component $z_\ell(n)$ we start by analyzing the autocorrelation of the reverberant signal $z(n)$. The autocorrelation of the reverberant signal $z(n)$ at discrete time n and lag τ for a fixed source-microphone configuration is defined as

$$r_{zz}(n,n+\tau;h) = E\{z(n)z(n+\tau)\}, \tag{3.65}$$

where $E\{\cdot\}$ denotes ensemble averaging. Using (3.37), we have for one realization of h,

$$\begin{aligned} r_{zz}(n,n+\tau;h) = &\sum_{l=n-n_{\mathrm{d}}+1}^{n} \sum_{l'=n-n_{\mathrm{d}}+1+\tau}^{n+\tau} E\{s(l)s(l')\}h_{\mathrm{d}}(n-l)h_{\mathrm{d}}(n+\tau-l') \\ &+ \sum_{l=-\infty}^{n-n_{\mathrm{d}}} \sum_{l'=-\infty}^{n-n_{\mathrm{d}}+\tau} E\{s(l)s(l')\}h_{\mathrm{r}}(n-l)h_{\mathrm{r}}(n+\tau-l'). \end{aligned} \tag{3.66}$$

Using (3.6)–(3.8) and the fact that $b_{\mathrm{d}}(n)$ and $b_{\mathrm{r}}(n)$ consist of a zero-mean white Gaussian noise sequence, it follows that

$$\mathscr{E}\{h_{\mathrm{d}}(n-l)h_{\mathrm{d}}(n+\tau-l')\} = \sigma_{\mathrm{d}}^2 \mathrm{e}^{-2\bar{\zeta}n} \mathrm{e}^{\bar{\zeta}(l+l'-\tau)} \delta(l-l'+\tau), \tag{3.67}$$

and

$$\mathscr{E}\{h_{\mathrm{r}}(n-l)h_{\mathrm{r}}(n+\tau-l')\} = \sigma_{\mathrm{r}}^2 \mathrm{e}^{-2\bar{\zeta}n} \mathrm{e}^{\bar{\zeta}(l+l'-\tau)} \delta(l-l'+\tau), \tag{3.68}$$

where $\delta(\cdot)$ denotes the Kronecker delta function. It should be noted that $\mathscr{E}\{b_{\mathrm{d}}(n)b_{\mathrm{r}}(n+\tau)\} = 0$ implies that $\mathscr{E}\{h_{\mathrm{d}}(n)h_{\mathrm{r}}(n+\tau)\} = 0$. Under the assumption that the stochastic processes h and s are mutually independent the spatially averaged autocorrelation results in

$$\begin{aligned} r_{zz}(n,n+\tau) &= \mathscr{E}\{r_{zz}(n,n+\tau;h)\} \\ &= r_{z_{\mathrm{d}}z_{\mathrm{d}}}(n,n+\tau) + r_{z_{\mathrm{r}}z_{\mathrm{r}}}(n,n+\tau), \end{aligned} \tag{3.69}$$

with

$$r_{z_{\mathrm{d}}z_{\mathrm{d}}}(n,n+\tau) = \mathrm{e}^{-2\bar{\zeta}n} \sum_{l=n-n_{\mathrm{d}}+1}^{n} E\{s(l)s(l+\tau)\}\, \sigma_{\mathrm{d}}^2 \mathrm{e}^{2\bar{\zeta}l}, \tag{3.70}$$

and

$$r_{z_\mathrm{r}z_\mathrm{r}}(n,n+\tau) = \mathrm{e}^{-2\bar{\zeta}n} \sum_{l=-\infty}^{n-n_\mathrm{d}} E\{s(l)s(l+\tau)\}\, \sigma_\mathrm{r}^2 \mathrm{e}^{2\bar{\zeta}l} \tag{3.71}$$

$$= \mathrm{e}^{-2\bar{\zeta}n} \sum_{l=n-2n_\mathrm{d}+1}^{n-n_\mathrm{d}} E\{s(l)s(l+\tau)\}\, \sigma_\mathrm{r}^2 \mathrm{e}^{2\bar{\zeta}l} + \mathrm{e}^{-2\bar{\zeta}n} \sum_{l=-\infty}^{n-2n_\mathrm{d}} E\{s(l)s(l+\tau)\}\, \sigma_\mathrm{r}^2 \mathrm{e}^{2\bar{\zeta}l}. \tag{3.72}$$

The first term in (3.69) depends on the direct signal between time $n-n_\mathrm{d}+1$ and n, and the second depends on the reverberant signal.

Let us consider the spatially averaged autocorrelation at time $n-n_\mathrm{d}$:

$$r_{zz}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau) = r_{z_\mathrm{d}z_\mathrm{d}}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau) + r_{z_\mathrm{r}z_\mathrm{r}}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau), \tag{3.73}$$

with

$$r_{z_\mathrm{d}z_\mathrm{d}}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau) = \sigma_\mathrm{d}^2\, \mathrm{e}^{-2\bar{\zeta}(n-n_\mathrm{d})} \sum_{l=n-2n_\mathrm{d}+1}^{n-n_\mathrm{d}} E\{s(l)s(l+\tau)\}\mathrm{e}^{2\bar{\zeta}l}, \tag{3.74}$$

and

$$r_{z_\mathrm{r}z_\mathrm{r}}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau) = \sigma_\mathrm{r}^2\, \mathrm{e}^{-2\bar{\zeta}(n-n_\mathrm{d})} \sum_{l=-\infty}^{n-2n_\mathrm{d}} E\{s(l)s(l+\tau)\}\mathrm{e}^{2\bar{\zeta}l}. \tag{3.75}$$

Using (3.74) and (3.75) the term $r_{z_\mathrm{r}z_\mathrm{r}}(n,n+\tau)$ can be expressed as

$$r_{z_\mathrm{r}z_\mathrm{r}}(n,n+\tau) = \kappa\, \mathrm{e}^{-2\bar{\zeta}n_\mathrm{d}} r_{z_\mathrm{d}z_\mathrm{d}}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau) + \mathrm{e}^{-2\bar{\zeta}n_\mathrm{d}}\, r_{z_\mathrm{r}z_\mathrm{r}}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau), \tag{3.76}$$

with $\kappa = \sigma_\mathrm{r}^2/\sigma_\mathrm{d}^2$. Here $\kappa \leq 1$, since it is assumed that $\sigma_\mathrm{d}^2 \geq \sigma_\mathrm{r}^2$. Using (3.73) we can rewrite (3.76) as

$$r_{z_\mathrm{r}z_\mathrm{r}}(n,n+\tau) = \mathrm{e}^{-2\bar{\zeta}n_\mathrm{d}}\,(1-\kappa)\, r_{z_\mathrm{r}z_\mathrm{r}}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau) + \kappa\, \mathrm{e}^{-2\bar{\zeta}n_\mathrm{d}}\, r_{zz}(n-n_\mathrm{d},n-n_\mathrm{d}+\tau). \tag{3.77}$$

The late reverberant component can now be obtained using

$$r_{z_\ell z_\ell}(n,n+\tau) = \mathrm{e}^{-2\bar{\zeta}(n_\mathrm{e}-n_\mathrm{d})} r_{z_\mathrm{r}z_\mathrm{r}}(n-n_\mathrm{e}+n_\mathrm{d},n-n_\mathrm{e}+n_\mathrm{d}+\tau). \tag{3.78}$$

Note that for $\kappa = 1$, i.e., $\sigma_\mathrm{d}^2 = \sigma_\mathrm{r}^2$, (3.77) and (3.78) result in

$$r_{z_\ell z_\ell}(n,n+\tau) = \mathrm{e}^{-2\bar{\zeta}n_\mathrm{e}} r_{zz}(n-n_\mathrm{e},n-n_\mathrm{e}+\tau). \tag{3.79}$$

Given an estimate of the reverberation time T_{60}, the parameter $\bar{\zeta}$ can be calculated using (3.4). The parameter $\kappa = \sigma_{\mathrm{r}}^2/\sigma_{\mathrm{d}}^2$ is related to the DRR, which is defined as

$$\frac{E_{\mathrm{d}}}{E_{\mathrm{r}}} = \frac{\sum_{l=0}^{n_{\mathrm{d}}} h^2(l)}{\sum_{l=n_{\mathrm{d}}+1}^{\infty} h^2(l)}. \tag{3.80}$$

It should be noted that the DRR can be estimated directly from the AIR using (3.80). However, in many practical situations the AIR is not known in advance. Therefore, we will discuss the blind estimation of the reverberation time T_{60} and κ in Section 3.7. Using the model in (3.6) the direct and reverberant energy can be expressed as

$$E_{\mathrm{d}} = \sum_{l=0}^{n_{\mathrm{d}}} \sigma_{\mathrm{d}}^2 \mathrm{e}^{-2\bar{\zeta}l} = \frac{\sigma_{\mathrm{d}}^2}{2\bar{\zeta}}\left(1 - \mathrm{e}^{-2\bar{\zeta}n_{\mathrm{d}}}\right) \tag{3.81}$$

and

$$E_{\mathrm{r}} = \sum_{l=n_{\mathrm{d}}+1}^{\infty} \sigma_{\mathrm{r}}^2 \mathrm{e}^{-2\bar{\zeta}l} = \frac{\sigma_{\mathrm{r}}^2}{2\bar{\zeta}} \mathrm{e}^{-2\bar{\zeta}n_{\mathrm{d}}}, \tag{3.82}$$

respectively, where σ_{d}^2 and σ_{r}^2 denote the variances of $b_{\mathrm{d}}(n)$ and $b_{\mathrm{r}}(n)$, respectively. Now the parameter κ can be expressed in terms of E_{d} and E_{r}:

$$\kappa = \frac{\sigma_{\mathrm{r}}^2}{\sigma_{\mathrm{d}}^2} = \frac{1 - \mathrm{e}^{-2\bar{\zeta}n_{\mathrm{d}}}}{\mathrm{e}^{-2\bar{\zeta}n_{\mathrm{d}}}} \frac{E_{\mathrm{r}}}{E_{\mathrm{d}}}. \tag{3.83}$$

In general the DRR is frequency dependent, as shown in [48]. Hence, to improve the accuracy of the model we propose to make κ frequency dependent. Furthermore, we should keep in mind that the DRR, and thus κ, depends on the distance between the source and microphone. Therefore, spatial averaging can only be performed over those microphone signals that have the same source-microphone distance.

In practice the signals can be considered as stationary over periods of time that are short compared to the reverberation time T_{60}. This is justified by the fact that the exponential decay is very slow and that speech is quasi-stationary. We consider that $n_{\mathrm{e}} \ll T_{60} f_{\mathrm{s}}$ and that $n_{\mathrm{e}}/f_{\mathrm{s}}$ is larger than the time span over which the speech signal can be considered stationary, which is usually around 20–40 ms [22]. In the following we assume that n_{d} is equal to the number of samples separating two successive STFT frames, denoted by R. Under these assumptions and by taking the frequency dependency of κ and $\bar{\zeta}$ into account, the counterparts of (3.77) and (3.78) in terms of the spectral variances are:

$$\lambda_{z_{\mathrm{r}}}(\ell,k) = \mathrm{e}^{-2\bar{\zeta}(k)R}\left(1 - \kappa(k)\right)\lambda_{z_{\mathrm{r}}}(\ell-1,k) + \kappa(k)\,\mathrm{e}^{-2\bar{\zeta}(k)R}\lambda_z(\ell-1,k), \tag{3.84}$$

and

$$\lambda_{z_\ell}(\ell,k) = \mathrm{e}^{-2\bar{\zeta}(k)(n_{\mathrm{e}}-R)}\lambda_{z_{\mathrm{r}}}(\ell - \frac{n_{\mathrm{e}}}{R} + 1, k). \tag{3.85}$$

Note that the value n_e should be chosen such that n_e/R is an integer value.

By substituting $\lambda_z(\ell,k) = \lambda_{z_d}(\ell,k) + \lambda_{z_r}(\ell,k)$ in (3.84) and rearranging the terms we obtain

$$\lambda_{z_r}(\ell,k) = \mathrm{e}^{-2\bar{\xi}(k)R}\lambda_{z_r}(\ell-1,k) + \kappa(k)\mathrm{e}^{-2\bar{\xi}(k)R}\lambda_{z_d}(\ell-1,k). \tag{3.86}$$

Using (3.83) we obtain

$$\lambda_{z_r}(\ell,k) = \mathrm{e}^{-2\bar{\xi}(k)R}\lambda_{z_r}(\ell-1,k) + \frac{E_r}{E_d}\left(1-\mathrm{e}^{-2\bar{\xi}(k)R}\right)\lambda_{z_d}(\ell-1,k). \tag{3.87}$$

This equation shows that the spectral variance of the reverberant signal component at time frame ℓ consists of $\mathrm{e}^{-2\bar{\xi}(k)R}$ times the spectral variance of the reverberant signal component at time frame $\ell-1$ and $\frac{E_r}{E_d}\left(1-\mathrm{e}^{-2\bar{\xi}(k)R}\right)$ times the spectral variance of the direct speech component at time frame $\ell-1$. While the first term models the energy decay in the room, the second term models the energy growth due to the power of the source ($\lambda_{z_d}(\ell,k)/E_d$). As expected, only the source can increase the reverberant energy in the room and the absorption of the energy is completely determined by the reverberation time of the room.

3.7 Estimating Model Parameters

In order to estimate the late reverberant spectral variance an estimate of the reverberation time T_{60} of the room and the direct to reverberation ratio is required.

3.7.1 Reverberation Time

Partially blind methods to estimate the reverberation time have been developed in which the characteristics of the room are learnt using neural network approaches [19]. Another method uses a segmentation procedure for detecting gaps in the signals and then tracks the sound decay curve [49, 74]. A blind method has been proposed by Ratnam *et al.* based on a maximum-likelihood estimation procedure [64]. In [53] Löllmann and Vary proposed a maximum-likelihood estimator which takes additive noise into account. Most of these methods can also be applied to band-pass filtered versions of the original signal in order to estimate the reverberation time in different frequency bands.

In general, it is reasonable to assume that the reverberation time is approximately constant in the room. Therefore, in communication systems that involve echo cancellation, the reverberation time can be estimated using the estimated echo path [41]. For some applications such as audio or video-conferencing where a fixed setup is used, the reverberation time can be estimated using a calibration process.

3.7.2 Direct-to-reverberant Ratio

In many practical situations the distance between the source and the microphone will vary. Since the DRR depends on the distance between the source and the microphone, it is important that the parameter κ can be estimated online.

The parameter κ was introduced to prevent over-estimation of the reverberant spectral variance $\lambda_{z_\mathrm{r}}(\ell,k)$ when the source-microphone distance is smaller than the critical distance. In the case when κ is too large, the spectral variance $\hat{\lambda}_{z_\mathrm{r}}(\ell,k)$ could become larger than $|Z(\ell,k)|^2$, which indicates that over-estimation has occurred. In this case, the value of κ should be lowered. In addition we know that during the free decay, which occurs after an offset of the source signal, $\hat{\lambda}_{z_\mathrm{r}}(\ell,k)$ should be approximately equal to $|Z(\ell,k)|^2$. Estimation of κ could therefore be performed after a speech offset. Unfortunately, the detection of speech offsets is rather difficult. However, from the above discussion is has become clear that κ should at least fulfill the following condition: $|Z(\ell,k)|^2 - \hat{\lambda}_{z_\mathrm{r}}(\ell,k) \geq 0$.

The parameter κ can be estimated adaptively using the following strategy: (1) when speech is detected and $|Z(\ell,k)|^2 < \hat{\lambda}_{z_\mathrm{r}}(\ell,k)$ the value of κ is lowered, (2) when $|Z(\ell,k)|^2 > \hat{\lambda}_{z_\mathrm{r}}(\ell,k)$ the value of κ is raised slowly and (3) when $|Z(\ell,k)|^2 = \hat{\lambda}_{z_\mathrm{r}}(\ell,k)$ the value of κ is assumed to be correct. This strategy can be implemented as follows:

$$\hat{\kappa}(\ell) = \hat{\kappa}(\ell-1) + \frac{\mu_\kappa}{P_z(\ell-1)} \sum_{k=0}^{\frac{K}{2}-1} \left(|Z(\ell-1,k)|^2 - \hat{\lambda}_{z_\mathrm{r}}(\ell-1,k) \right) \tag{3.88}$$

where $P_z(\ell-1) = \sum_{k=0}^{\frac{K}{2}-1} |Z(\ell-1,k)|^2$, and μ_κ ($0 < \mu_\kappa < 1$) denotes the step-size. After each update step, $\hat{\kappa}(\ell)$ is constrained, such that $0 < \hat{\kappa}(\ell) \leq 1$. Experimental results that demonstrate the feasibility of this estimator can be found in Sect. 3.8.

3.8 Experimental Results

In this section we present and discuss the experimental results that were obtained using single and multiple microphones. A uniformly linear microphone array was used with inter-microphone spacing $D_\mathrm{i} = 5$ cm. The source-array distance D is defined as the distance between the source and the center of the array, and ranges from 0.25 to 3 m. The dimensions of the room are 5 m $\times$ 6 m $\times$ 4 m (length $\times$ width $\times$ height). The experimental setup is depicted in Fig. 3.3. The APLAWD database [51] was used for evaluation with the sampling frequency set to $f_\mathrm{s} = 8$ kHz; it contains anechoic recordings comprising ten repetitions of five sentences uttered by five male and five female talkers. The reverberant microphone signals were obtained by convolving the anechoic recordings with different AIRs. The AIRs are generated using the image method for modelling small room acoustics [5], modified to accommodate fractional sample delays according to [59], with reverberation times from 250 to 1000 ms. The additive noise $v(n)$ was speech-like noise, taken

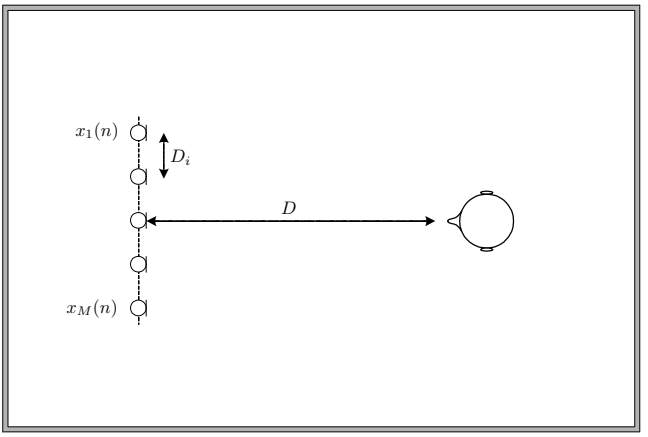


Fig. 3.3 Experimental setup with a uniform linear microphone array

from the NOISEX-92 database [75]. The spectral variance of the noise was estimated from the noisy microphone signal $x(n)$ using the IMCRA approach [14]. All *a priori* SIRs were estimated using the decision-directed approach. In all experiments we assumed that the reverberation time T_{60} of the room is known. Its value was determined using the Schroeder method, described in [68]. The parameter κ was estimated adaptively using the method described in Sect. 3.7.2. The parameters that were used for these experiments are shown in Table 3.1.

The segmental SIR and Bark Spectral Distortion (BSD), as defined in Chap. 2, are used for the evaluation.

Table 3.1 Parameters used in experiments

$f_s = 8000$ Hz	$n_e = 40$ ms	$G_{\min}^{\mathrm{dB}} = 18$ dB	$\beta^{\mathrm{dB}} = 3$ dB
$\eta = 0.95$	b = Hanning window	$w_{\mathrm{MSC}} = 9$	$\Phi_{\min} = 0.2$
$\Phi_{\max} = 0.65$	$\eta_s = 0.35$		

3.8.1 Using One Microphone

In this section we evaluate the performance of the single-microphone dereverberation method in the presence of noise using two objective measures. A summary of the complete single-microphone spectral enhancement algorithm that suppresses late reverberation and ambient noise is summarized in Algorithm 3.1.

We first evaluate the objective measures when $T_{60} = 0.5$ s and $D = 1$ m. The Signal to Noise Ratio (SNR) of the microphone signal ranges from 10 to 30 dB. In

Algorithm 3.1 Summary of the single-microphone spectral enhancement algorithm that suppresses late reverberation and ambient noise

1. **STFT**: Calculate the STFT of the noisy and reverberant signal $x(n)$.
2. **Estimate model parameters**: Firstly, decay-rate $\bar{\zeta}(k)$ is calculated using (3.4). Secondly, the parameter κ is estimated using (3.88).
3. **Estimate ambient noise**: Estimate $\lambda_v(\ell,k)$ using the method described in [18].
4. **Estimate late reverberant energy**: Calculate $G_{\mathrm{SP}}(\ell,k)$ using (3.62)–(3.64). Estimate $\lambda_z(\ell,k)$ using (3.61), and calculate $\hat{\lambda}_{z_\ell}(\ell,k)$ using (3.85).
5. **Post-filter**:
   (a) Calculate the *a posteriori* SIR using (3.26) and the individual *a priori* SIRs using (3.35)–(3.36) with $\vartheta \in \{z_\ell, v\}$, the total *a priori* SIR can then be calculated using (3.33).
   (b) Estimate the *a priori* speech presence probability $p(\ell,k|\ell-1)$ using the method described in [15] and calculate $\hat{p}(\ell,k)$ using (3.54).
   (c) Calculate the gain function $G_{\text{MMSE-LSA}}(\ell,k)$ using (3.27), (3.29), and (3.24).
   (d) Calculate $\hat{Z}_{\mathrm{e}}(\ell,k)$ using (3.30).
6. **Inverse STFT**: Calculate the output $\hat{z}_{\mathrm{e}}(n)$ by applying the inverse STFT to $\hat{Z}_{\mathrm{e}}(\ell,k)$.

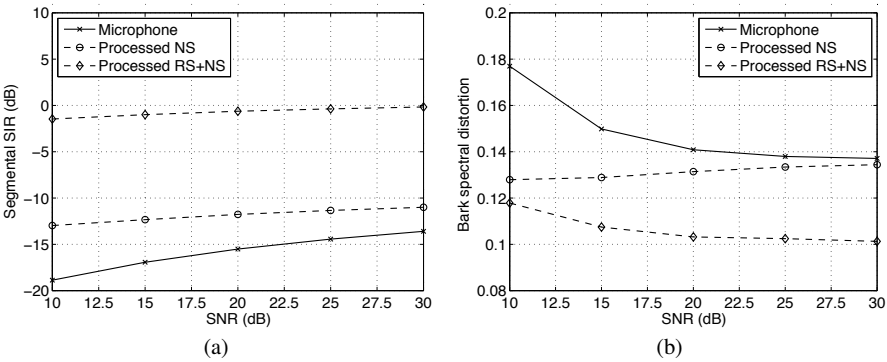


Fig. 3.4 (a) Segmental SIRs and (b) BSDs of the unprocessed microphone signal, the processed signal after noise suppression (NS), and the processed signal after joint reverberation and noise suppression (RS+NS). The SNR of the received signal varies between 10 and 30 dB ($D = 1$ m,$T_{60} =$ 500 ms, and $n_{\mathrm{e}}/f_{\mathrm{s}} = 40$ ms)

Fig. 3.4 the segmental SIR and BSD are depicted for the (unprocessed) reverberant microphone signal, the signal that was obtained after noise suppression (NS), and the signal that was obtained after joint reverberation and noise suppression (RS+NS). Joint reverberant and noise suppression significantly improves the segmental SIR (approximately 10 dB) and the BSD (approximately 0.04–0.06) compared to noise suppression only. After the noise suppression is applied, the reverberation becomes more pronounced. When, in addition to the noise, the late reverberation is suppressed, the subjective sound quality is significantly improved and the residual ambient noise sounds stationary. When listening to the processed signal, minor artifacts were audible when the SNR was larger than 15 dB. In Fig. 3.5,

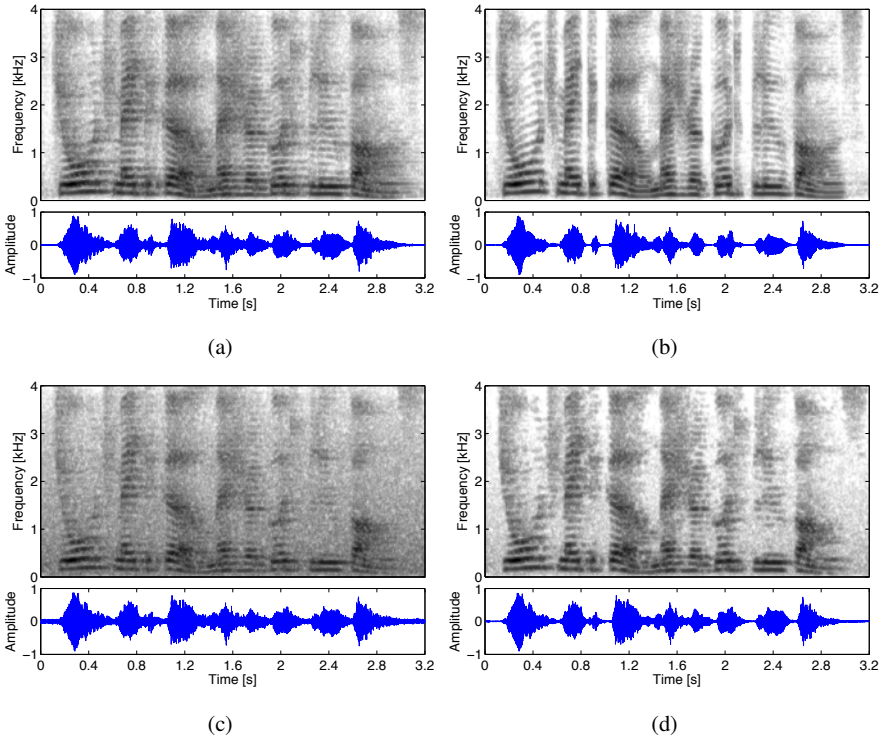
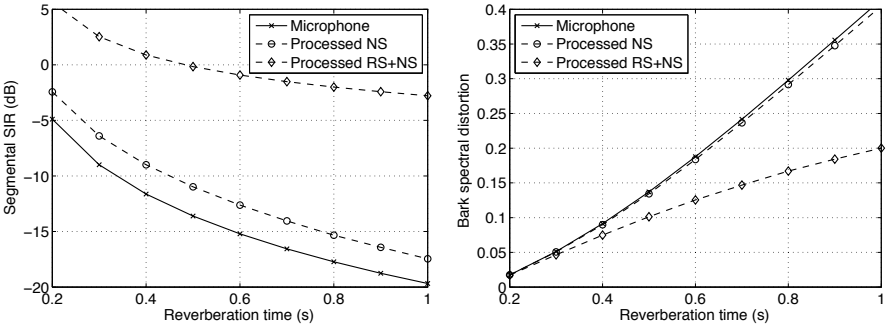


Fig. 3.5 Spectrograms and time-domain waveforms of (a) reverberant signal $z(n)$, (b) early speech signal $z_{\mathrm{e}}(n)$, (c) microphone signal (SNR $= 15$ dB, $T_{60} = 0.5$ s, $D = 1$ m), and (d) estimated early speech signal $\hat{z}_{\mathrm{e}}(n)$

spectrograms and time-domain waveforms are presented for one speech fragment. In both the spectrogram and time-domain waveform of the reverberant signal smearing of the speech, caused by the late reflections can be observed. In the enhanced speech signal, the smearing is significantly reduced as a result of the suppression of late reverberation. In addition, it can be seen that the noise is suppressed.

In the second experiment we evaluate the algorithms for SNR $= 30$ dB and $D =$ 1 m. The reverberation time T_{60} ranges from 0.2 to 1 s. In Fig. 3.6 the segmental SIR and BSD are depicted for the reverberant microphone signal, the signal that was obtained after noise suppression (NS), and the signal that was obtained after joint reverberation and noise suppression (RS+NS). Since the SNR is relatively high, the segmental SIR mainly depends on the reverberation suppression. The results of this experiment demonstrate that the algorithm is able to suppress a significant amount of late reverberation for short and long reverberation times. The results of an informal listening test indicated that for long reverberation times ($T_{60} > 0.5$ s), a larger value of n_{e} is preferred to maintain a natural sounding speech signal.



(a) (b)

Fig. 3.6 (a) Segmental SIRs and (b) BSDs of the unprocessed microphone signal, the processed signal after noise suppression (NS), and the processed signal after joint reverberation and noise suppression (RS+NS). The reverberation time varies between 0.2 and 1 s (SNR = 30 dB, $D = 1$ m, and $n_e/f_s = 40$ ms)

In the third experiment we evaluate the algorithms for SNR = 30 dB and $T_{60} = 0.5$ s. The source-microphone distance D ranges from 0.25 to 4 m. In the current setup the critical distance D_c equals 0.9 m. In Fig. 3.7 the segmental SIR and BSD are depicted for the reverberant microphone signal, the signal that was obtained after noise suppression (NS), and the signal that was obtained after joint reverberation and noise suppression (RS+NS). Since the SNR is relatively high, the segmental SIR mainly depends on the reverberation suppression. The results shown here demonstrate that the algorithm is able to suppress a significant amount of late reverberation over a wide range of source-microphone distances that are smaller and larger than the critical distance. While the BSD measures mainly show an improvement when the source-microphone distances are large, the segmental SIR improvement is almost constant. It should be noted that, for a source-microphone distance smaller than the critical distance, the value of n_e/f_s can be decreased without affecting the amount of speech distortion significantly.

3.8.2 Using Multiple Microphones

In this section we evaluate the performance of three multi-microphone dereverberation methods in the presence of spatially white noise (SNR = 30 dB) using two objective measures. Since the SNR is relatively high, the segmental SIR mainly depends on the reverberation suppression. The first multi-microphone method is the Delay-and-sum Beamformer (DSB). The second method is the delay and sum beamformer in conjunction with the single-channel post-filter described in Algorithm. 3.1 and is denoted by (DSB-PF). The third method is based on the non-linear spatial processor in conjunction with the same single-channel post-filter and is denoted by

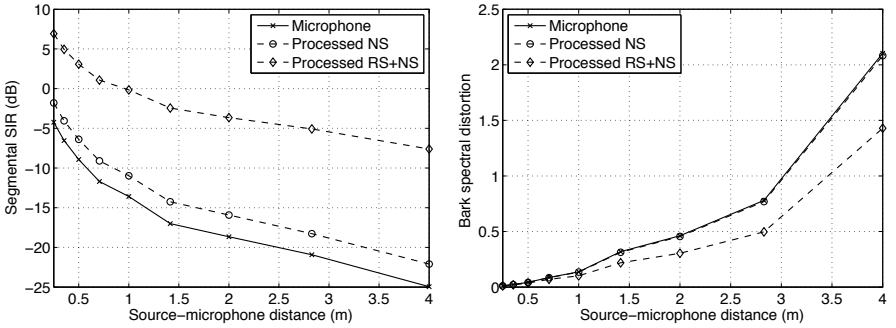
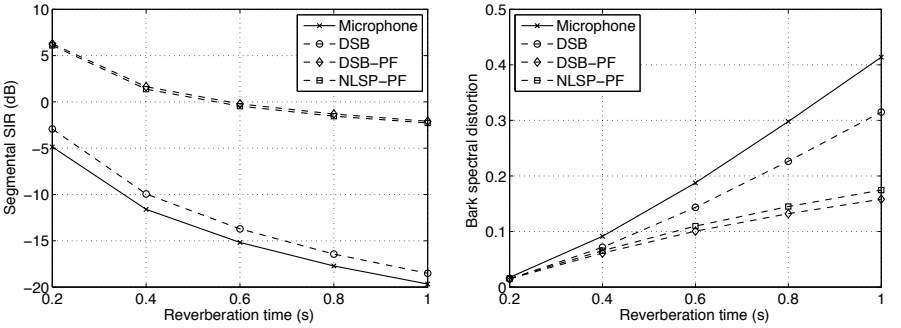


Fig. 3.7 (a) Segmental SIRs and (b) BSDs of the unprocessed microphone signal, the processed signal after noise suppression (NS), and the processed signal after joint reverberation and noise suppression (RS+NS). The source-microphone varies between 0.25 and 4 m (SNR = 30 dB, T_{60} = 500 ms, and $n_e/f_s = 40$ ms)

(NLSP-PF). As a reference the signal of the microphone that is closest to the desired source was evaluated.

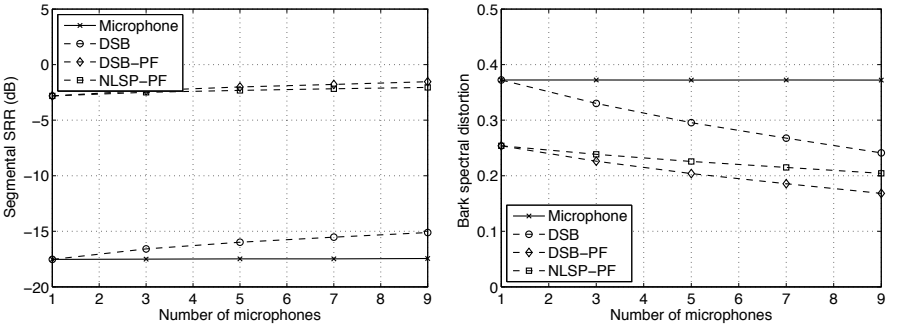
In the first experiment the number of microphones used was $M = 5$ and the source-microphone distance was set to $D = 1.5$ m. The reverberation time T_{60} ranged from 0.2 to 1 s. In Fig. 3.8 the segmental SIR and BSD are depicted for the reference microphone signal, the output of the DSB, the result of the DSB-PF method, and the result of the NLSP-PF method. These results show the limited performance of the DSB. A significant improvement is achieved by applying the single-channel post-filter to the output of the delay and sum beamformer. According to the objective measures employed the NLSP-PF method performs slightly worse compared to the DSB-PF method. However, the results of an informal listing test indicated that the output of the NLSP-PF method sounds more natural and contains less audible distortions than the output of the DSB-PF method. This could be explained by the fact that the objective measures used in this work are unable to reflect certain perceptual characteristics of the evaluated signals that are important in the context of speech dereverberation.

In the second experiment the reverberation time $T_{60} = 0.5$ s was used, and the source-microphone distance was set to $D = 1.5$ m. The number of microphones M ranged from 1 to 9. The segmental SIR and BSD values obtained are shown in Fig. 3.9. As in the previous experiment we can see that the single-channel post-filter significantly increases the dereverberation performance. The segmental SIR was increased by more than 14.5 dB compared to the reference microphone. It is noted that the segmental SIR increases slightly when more than one microphone is used. However, the BSD is significantly reduced by using multi-microphone signals. In terms of the segmental SIR and BSD the best result is obtained by the DSB-PF system. Judging from these results one might argue that the DSB-PF method performs better than the NLSP-PF method. However, as before the results from an



(a) (b)

Fig. 3.8 (a) Segmental SIRs and (b) BSDs of the reference microphone signal, the DSB signal, the DSB-PF signal, and the NLSP-PF signal. The reverberation time varies between 0.2 and 1 s ($D = 1.5$ m, SNR $= 30$ dB, and $n_e/f_s = 40$ ms)



(a) (b)

Fig. 3.9 (a) Segmental SIRs and (b) BSDs of the reference microphone signal, the DSB signal, the DSB-PF signal, and the NLSP-PF signal. The number of microphones ranges from 1 to 9 ($D = 1.5$ m, $T_{60} = 0.5$ s, SNR $= 30$ dB, and $n_e/f_s = 40$ ms)

informal listening test indicated that the results obtained by the NLSP-PF method sound more natural and contain less artifacts than the results obtained by the DSB-PF method.

3.9 Summary and Outlook

In this chapter single and multi-microphone speech dereverberation methods that are entirely or partly based on spectral enhancement were described. The quality of the received speech signal can be improved by reducing the effective noise that consists

of late reverberation and ambient noise. It was shown that quantifiable properties of the AIR, such as the reverberation time and DRR, can be used to dereverberate the received speech signal partly. In order to use spectral enhancement methods for speech dereverberation, an estimate of the late reverberant spectral variance is required. In Sect. 3.6 such an estimator was derived using a generalized statistical reverberation model. When the source-receiver distance is smaller than the critical distance the proposed estimator that is based on the generalized statistical model is advantageous over the estimator that is based on Polack's statistical model [39].

In the development of the speech enhancement method we assumed that the spectral coefficients of the speech and noise are Gaussian. Furthermore, we used the minimum mean squared error distortion measure and the log-amplitude fidelity criterion that was successfully used for noise suppression. However, it has yet to be determined if the MSE distortion measure and log-amplitude fidelity criterion provide the best results in the case of reverberation and noise suppression. Recently, the generalized autoregressive conditional heteroscedasticity (GARCH) model was shown to be useful for statistically modelling speech signals in the STFT domain [16]. A Markov-switching time-frequency GARCH model was proposed in [1, 2] for modelling non-stationary signals in the time-frequency domain. The model takes into account the strong correlation of successive spectral magnitudes and is more appropriate than the decision-directed approach for speech spectral variance estimation in noisy environments. Should this or other statistical speech models be used in the development of novel spectral speech dereverberation algorithms, they might further increase the suppression of late reverberation and noise and decrease the amount of speech distortion. In the course of this chapter, two modifications of the standard MMSE-LSA estimator were discussed. The first modification concerns the spectral gain function and allows a larger suppression of late reverberation when the early speech component is inactive and results in a constant residual ambient noise level. The second modification concerns the speech presence probability estimator, which is improved by analyzing the magnitude squared coherence of the observed sound field.

We also investigated the use of multiple microphones for speech dereverberation and described two multi-microphone systems. The first system consists of an MVDR beamformer followed by a single-channel post-filter. Although this system can be useful in the presence of coherent noise sources, we could not directly exploit the spatial diversity of the reverberant signal to estimate the late reverberant spectral variance. In a spatially white noise field, the MVDR beamformer reduces to the well-known delay and sum beamformer. It has been shown in [39] that due to the spatial correlation between the AIRs, the residual reverberation at the output of the beamformer might contain undesired signal components. These components are especially pronounced at low frequencies and become larger when the inter-microphone distances are small. A second multi-microphone system that does not suffer from the spatial correlation between the AIRs was described. The latter consists of a non-linear spatial processor followed by a single-channel post-filter. The non-linear spatial processor can only be employed when the noise field is spatially white. Although practically feasible multi-microphone solutions have been found,

further research is required to investigate the tradeoff between noise suppression and reverberation suppression.

Finally, experimental results demonstrated the beneficial use of the single-microphone spectral dereverberation method described and showed that a large amount of reverberation and noise can be reduced with little speech distortion.

Acknowledgment

The author thanks Dr. Sharon Gannot and Dr. Israel Cohen for the valuable discussions and helpful suggestions.

References

1. Abramson, A., Cohen, I.: Markov-switching GARCH model and application to speech enhancement in subbands. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 1–4. Paris, France (2006)
2. Abramson, A., Cohen, I.: Recursive supervised estimation of a Markov-switching GARCH process in the short-time Fourier transform domain. IEEE Trans. Signal Process. **55**(7), 3227–3238 (2007)
3. Accardi, A.J., Cox, R.V.: A modular approach to speech enhancement with an application to speech coding. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 201–204 (1999)
4. Allen, J.B.: Effects of small room reverberation on subjective preference. J. Acoust. Soc. Am. **71**(S1), S5 (1982)
5. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
6. Benesty, J., Makino, S., Chen, J. (eds.): Speech Enhancement. Springer (2005)
7. Benesty, J., Sondhi, M.M., Huang, Y. (eds.): Springer Handbook of Speech Processing. Springer (2007)
8. Berouti, M., Schwartz, R., Makhoul, J.: Enhancement of speech corrupted by acoustic noise. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 4, pp. 208–211 (1979)
9. Boll, S.F.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust., Speech, Signal Process. **ASSP-27**(2), 113–120 (1979)
10. Bolt, R.H., MacDonald, A.D.: Theory of speech masking by reverberation. J. Acoust. Soc. Am. **21**, 577–580 (1949)
11. Burshtein, D., Gannot, S.: Speech enhancement using a mixture-maximum model. IEEE Trans. Speech Audio Process. **10**(6), 341351 (2002)
12. Cappe, O.: Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech Audio Process. **2**(2), 345–349 (1994). DOI 10.1109/89.279283
13. Cohen, I.: Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. IEEE Signal Process. Lett. **9**(4), 113–116 (2002)
14. Cohen, I.: Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. **11**(5), 466–475 (2003). DOI 10.1109/TSA.2003.811544

15. Cohen, I.: From volatility modeling of financial time-series to stochastic modeling and enhancement of speech signals. In: J. Benesty, S. Makino, J. Chen (eds.) Speech Enhancement, chap. 5, pp. 97–114. Springer (2005)
16. Cohen, I.: Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models. Signal Processing **86**(4), 698–709 (2006)
17. Cohen, I., Gannot, S.: Spectral enhancement methods. In: Benesty et al. [7], chap. 45. Part H
18. Cohen, I., Gannot, S., Berdugo, B.: An integrated real-time beamforming and post filtering system for nonstationary noise environments. EURASIP J. on App. Signal Process. **11**, 1064–1073 (2003)
19. Cox, T.J., Li, F., Darlington, P.: Extracting room reverberation time from speech using artificial neural networks. J. Audio Eng. Soc. **49**(4), 219–230 (2001)
20. Crochiere, R.E., Rabiner, L.R.: Multirate Digital Signal Processing. Prentice-Hall (1983)
21. Delcroix, M., Hikichi, T., Miyoshi, M.: Precise dereverberation using multichannel linear prediction. IEEE Trans. Audio, Speech, Lang. Process. **15**(2), 430–440 (2007)
22. Deller, J.R., Proakis, J.G., Hansen, J.H.L.: Discrete-Time Processing of Speech Signals. New York: MacMillan (1993)
23. Ephraim, Y., Cohen, I.: Recent advancements in speech enhancement. In: R.C. Dorf (ed.) The Electrical Engineering Handbook, Circuits, Signals, and Speech and Image Processing, third edn. CRC Press (2006)
24. Ephraim, Y., Lev-Ari, H., Roberts, W.J.J.: A brief survey of speech enhancement. In: The Electronic Handbook, second edn. CRC Press (2005)
25. Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process. **32**(6), 1109–1121 (1984)
26. Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process. **33**(2), 443–445 (1985)
27. Gannot, S., Cohen, I.: Adaptive beamforming and postfiltering. In: Benesty et al. [7], chap. 48
28. Gannot, S., Moonen, M.: Subspace methods for multimicrophone speech dereverberation. EURASIP J. on App. Signal Process. **2003**(11), 1074–1090 (2003)
29. Gaubitch, N.D., Naylor, P.A.: Analysis of the dereverberation performance of microphone arrays. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC) (2005)
30. Gaubitch, N.D., Naylor, P.A., Ward, D.B.: On the use of linear prediction for dereverberation of speech. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 99–102 (2003)
31. Goh, Z., Tan, K.C., Tan, T.G.: Postprocessing method for suppressing musical noise generated by spectral subtraction. IEEE Trans. Speech Audio Process. **6**(3), 287–292 (1998). DOI 10.1109/89.668822
32. Griebel, S.M., Brandstein, M.S.: Wavelet transform extrema clustering for multi-channel speech dereverberation. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 52–55. Pocono Manor, Pennsylvania (1999)
33. Gürelli, M.I., Nikias, C.L.: EVAM: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals. IEEE Trans. Signal Process. **43**(1), 134–149 (1995)
34. Gustafsson, S., Martin, R., Jax, P., Vary, P.: A psychoacoustic approach to combined acoustic echo cancellation and noise reduction. IEEE Trans. Speech Audio Process. **10**(5), 245–256 (2002)
35. Gustafsson, S., Martin, R., Vary, P.: Combined acoustic echo control and noise reduction for hands-free telephony. Signal Processing **64**(1), 21–32 (1998)
36. Gustafsson, S., Nordholm, S., Claesson, I.: Spectral subtraction using reduced delay convolution and adaptive averaging. IEEE Trans. Speech Audio Process. **9**(8), 799–807 (2001)
37. Habets, E.A.P.: Multi-channel speech dereverberation based on a statistical model of late reverberation. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 173–176. Philadelphia, USA (2005)
38. Habets, E.A.P.: Speech dereverberation based on a statistical model of late reverberation using a linear microphone array. In: Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA), pp. d7–d8. Piscataway, NJ, USA (2005)

39. Habets, E.A.P.: Single- and multi-microphone speech dereverberation using spectral enhancement. Ph.D. thesis, Technische Universiteit Eindhoven (2007)
40. Habets, E.A.P., Cohen, I., Gannot, S.: MMSE log spectral amplitude estimator for multiple interferences. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 1–4. Paris, France (2006)
41. Habets, E.A.P., Cohen, I., Gannot, S., Sommen, P.C.W.: Joint dereverberation and residual echo suppression of speech signals in noisy environments. IEEE Trans. Audio, Speech, Lang. Process. **16**(8), 1433–1451 (2008)
42. Habets, E.A.P., Gannot, S., Cohen, I.: Dual-microphone speech dereverberation in a noisy environment. In: Proc. IEEE Int. Symposium on Signal Processing and Information Technology (ISSPIT), pp. 651–655. Vancouver, Canada (2006)
43. Haykin, S.: Blind Deconvolution, fourth edn. Prentice-Hall Information and System Sciences. Prentice-Hall (1994)
44. Hopgood, J.: Nonstationary signal processing with application to reverberation cancellation in acoustic environments. Ph.D. thesis, Cambridge University (2001)
45. Huang, Y., Benesty, J.: A class of frequency-domain adaptive approaches to blind multichannel identification. IEEE Trans. Signal Process. **51**(1), 11–24 (2003)
46. Jetzt, J.J.: Critical distance measurement of rooms from the sound energy spectral response. J. Acoust. Soc. Am. **65**(5), 1204–1211 (1979)
47. Jot, J.M., Cerveau, L., Warusfel, O.: Analysis and synthesis of room reverberation based on a statistical time-frequency model. In: Proc. Audio Eng. Soc. Convention (1997)
48. Kuttruff, H.: Room Acoustics, 4th edn. Taylor & Frances (2000)
49. Lebart, K., Boucher, J.M., Denbigh, P.N.: A new method based on spectral subtraction for speech dereverberation. Acta Acoustica **87**, 359–366 (2001)
50. Lim, J.S., Oppenheim, A.V.: Enhancement and bandwidth compression of noisy speech. Proc. IEEE **67**(12), 1586–1604 (1979)
51. Lindsey, G., Breen, A., Nevard, S.: SPAR's archivable actual-word databases. Technical report, University College London (1987)
52. Loizou, P.C.: Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum. IEEE Trans. Speech Audio Process. **13**(5), 857–869 (2005). DOI 10.1109/TSA.2005.851929
53. Löllmann, H.W., Vary, P.: Estimation of the reverberation time in noisy environments. In: Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC), pp. 1–4 (2008)
54. Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**, 504–512 (2001). DOI 10.1109/89.928915
55. Martin, R.: Speech enhancement based on minimum mean-square error estimation and supergaussian priors. IEEE Trans. Speech Audio Process. **13**(5), 845–856 (2005). DOI 10.1109/TSA.2005.851927
56. Miyoshi, M., Kaneda, Y.: Inverse filtering of room acoustics. IEEE Trans. Acoust., Speech, Signal Process. **36**(2), 145–152 (1988)
57. Nábělek, A.K., Letowski, T.R., Tucker, F.M.: Reverberant overlap- and self-masking in consonant identification. J. Acoust. Soc. Am. **86**(4), 1259–1265 (1989)
58. Nábělek, A.K., Mason, D.: Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms. J. Speech Hear. Res. **24**, 375–383 (1981)
59. Peterson, P.M.: Simulating the response of multiple microphones to a single acoustic source in a reverberant room. J. Acoust. Soc. Am. **80**(5), 1527–1529 (1986)
60. Peutz, V.M.A.: Articulation loss of consonants as a criterion for speech transmission in a room. J. Audio Eng. Soc. **19**(11), 915–919 (1971)
61. Polack, J.D.: La transmission de l'énergie sonore dans les salles. Ph.D. thesis, Université du Maine, La Mans, France (1988)
62. Polack, J.D.: Playing billiards in the concert hall: the mathematical foundations of geometrical room acoustics. Appl. Acoust. **38**(2), 235–244 (1993)
63. Radlović, B.D., Kennedy, R.A.: Nonminimum-phase equalization and its subjective importance in room acoustics. IEEE Trans. Speech Audio Process. **8**(6), 728–737 (2000)

64. Ratnam, R., Jones, D.L., Wheeler, B.C., O'Brien, Jr., W.D., Lansing, C.R., Feng, A.S.: Blind estimation of reverberation time. J. Acoust. Soc. Am. **114**(5), 2877–2892 (2003)
65. Sabine, W.C.: Collected Papers on acoustics (Originally 1921). Peninsula Publishing (1993)
66. Schroeder, M.R.: Statistical parameters of the frequency response curves of large rooms. J. Audio Eng. Soc. **35**, 299–306 (1954)
67. Schroeder, M.R.: Frequency correlation functions of frequency responses in rooms. J. Acoust. Soc. Am. **34**(12), 1819–1823 (1962)
68. Schroeder, M.R.: Integrated-impulse method measuring sound decay without using impulses. J. Acoust. Soc. Am. **66**(2), 497–500 (1979)
69. Schroeder, M.R.: The "schroeder frequency" revisited. J. Acoust. Soc. Am. **99**(5), 3240–3241 (1996). DOI 10.1121/1.414868
70. Sim, B.L., Tong, Y.C., Chang, J.S., Tan, C.T.: A parametric formulation of the generalized spectral subtraction method. IEEE Trans. Speech Audio Process. **6**(4), 328–337 (1998)
71. Steinberg, J.C.: Effects of distortion upon the recognition of speech sounds. J. Acoust. Soc. Am. **1**, 35–35 (1929)
72. Takata, Y., Nábělek, A.K.: English consonant recognition in noise and in reverberation by Japanese and American listeners. J. Acoust. Soc. Am. **88**, 663–666 (1990)
73. Talantzis, F., Ward, D.B.: Robustness of multichannel equalization in an acoustic reverberant environment. J. Acoust. Soc. Am. **114**(2), 833–841 (2003)
74. Tashev, I., Malvar, H.S.: A new beamformer design algorithm for microphone arrays. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 3, pp. iii/101–iii/104 (2005)
75. Varga, A., Steeneken, H.J.M.: Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication **3**(3), 247–251 (1993). DOI 10.1016/0167--6393(93)90095--3
76. Wexler, J., Raz, S.: Discrete Gabor expansions. Signal Processing **21**(3), 207–220 (1990)
77. Wolfe, P.J., Godsill, S.J.: Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. EURASIP J. on App. Signal Process. **2003**(10), 1043–1051 (2003)
78. Yegnanarayana, B., Satyanarayana, P.: Enhancement of reverberant speech using LP residual signal. IEEE Trans. Speech Audio Process. **8**(3), 267–281 (2000)