Thesis for the degree of Doctor of Philosophy

Knowledge-Based Speech Enhancement

Sriram Srinivasan

KTH Electrical Engineering

Sound and Image Processing Laboratory
Department of Signals, Sensors and Systems
School of Electrical Engineering
KTH - Royal Institute of Technology

Stockholm 2005

Srinivasan, Sriram
Knowledge-Based Speech Enhancement

Copyright ©2005 Sriram Srinivasan except where otherwise stated. All rights reserved.

ISBN 91-628-6643-5
TRITA-S3-SIP 2005:1
ISSN 1652-4500
ISRN KTH/S3/SIP-05:01-SE

Sound and Image Processing Laboratory
Department of Signals, Sensors and Systems
KTH - Royal Institute of Technology
SE-100 44 Stockholm, Sweden
Tel: +46 8 790 7420

This thesis has been prepared using LaTeX.

Abstract

Speech is a fundamental means of human communication. In the last several decades, much effort has been devoted to the efficient transmission and storage of speech signals. With advances in technology making mobile communication ubiquitous, *communications anywhere* has become a reality. The freedom and flexibility offered by mobile technology brings with it new challenges, one of which is robustness to acoustic background noise. Speech enhancement systems form a vital front-end for mobile telephony in noisy environments such as in cars, cafeterias, subway stations, etc., in hearing aids, and to improve the performance of speech recognition systems.

In this thesis, which consists of four research articles, we discuss both single and multi-microphone approaches to speech enhancement. The main contribution of this thesis is a framework to exploit available prior knowledge about both speech and noise. The physiology of speech production places a constraint on the possible shapes of the speech spectral envelope, and this information is captured using codebooks of speech linear predictive (LP) coefficients obtained from a large training database. Similarly, information about commonly occurring noise types is captured using a set of noise codebooks, which can be combined with sound environment classification to treat different environments differently.

In paper A, we introduce maximum-likelihood estimation of the speech and noise LP parameters using the codebooks. The codebooks capture only the spectral shape. The speech and noise gain factors are obtained through a frame-by-frame optimization, providing good performance in practical nonstationary noise environments. The estimated parameters are subsequently used in a Wiener filter. Paper B describes Bayesian minimum mean squared error estimation of the speech and noise LP parameters and functions there-of, while retaining the instantaneous gain computation. Both memoryless and memory-based estimators are derived.

While papers A and B describe single-channel techniques, paper C describes a multi-channel Bayesian speech enhancement approach, where, in addition to temporal processing, the spatial diversity provided by multiple microphones is also exploited. In paper D, we introduce a multi-channel noise reduction technique motivated by blind source separation (BSS) concepts. In contrast to standard BSS approaches, we use the knowledge that one of the signals is speech and that the other is noise, and exploit their different characteristics.

Keywords: speech enhancement, noise reduction, linear predictive coefficients, autoregressive, codebooks, maximum-likelihood, Bayesian, nonstationary noise, blind source separation.

List of Papers

The thesis is based on the following papers:

[A] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, accepted for publication.

[B] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Transactions on Speech and Audio Processing*, submitted.

[C] S. Srinivasan, R. Aichner, W. B. Kleijn and W. Kellermann, "Multichannel parametric speech enhancement," *IEEE Signal Processing Letters*, accepted for publication.

[D] S. Srinivasan, M. Nilsson and W. B. Kleijn, "Speech denoising through source separation and min-max tracking," *IEEE Signal Processing Letters*, submitted.

In addition to papers A-D, the following refereed papers have also been published during the course of the PhD study:

[1] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Speech enhancement using a-priori information," in *Proceedings of Eurospeech*, vol. 2, Sept. 2003, pp. 1405–1408.

[2] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Estimation of short-term predictor parameters for coding and enhancement of noisy speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, May 2004, pp. 705–708.

[3] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Speech enhancement using a-priori information with classified noise codebooks," in *Proceedings XII European Signal Processing Conference*, Sept. 2004, pp. 1461–1464.

[4] V. Grancharov, S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Robust spectrum quantization for LP parameter enhancement," in *Proceedings XII European Signal Processing Conference*, Sept. 2004, pp. 1951-1954.

[5] S. Srinivasan and W. B. Kleijn, "Speech enhancement using adaptive time-domain segmentation," in *Proceedings of the International Conference on Spoken Language Processing*, vol. 2, Oct. 2004, pp. 869–872.

[6] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Mar. 2005, pp. 1077–1080.

[7] S. Srinivasan, M. Nilsson, and W. B. Kleijn, "Denoising through source separation and minimum tracking," in *Proceedings of Interspeech 2005 – Eurospeech*, Sept. 2005, pp. 2349–2352.

Acknowledgements

Several people have contributed to making the journey towards my Ph.D a memorable one. It is with great pleasure that I take this opportunity to acknowledge the support I have received.

First, I would like to express my gratitude to my supervisor, Prof. Bastiaan Kleijn, for his excellent guidance throughout my study. I have always admired your professionalism, enthusiasm, attention to detail, and sense of commitment. I have learnt a lot from you both in and out of work, and am thankful for all your help and advice.

My thanks to Dr. Jonas Samuelsson, my 'daily' supervisor, cannot be overstated. It was an absolute pleasure working with you, and I have benefited immensely from the innumerable discussions we have had. Your patience and pedagogic abilities have never ceased to inspire me. Thank you.

I would like to thank Prof. Arne Leijon for the interesting discussions and cooperation in the Hearcom project. To all my current and past colleagues at the Sound and Image Processing Laboratory: Anders, Barbara, David, Elisabet, Harald, Jan, Jesus, Jonas L, Kahye, Mattias, Moo Young, Renat, Volodya and Yusuke - thank you for making the Lab the fun-filled place that it is. Thanks are also due to the Lab's visitors, Prof. Peter Kabal and Dr. Geoffrey Chan, for stimulating discussions and ideas.

I thank Prof. Walter Kellermann and Robert Aichner for arranging the exchange visit at the Multimedia Communications and Signal Processing Group, University of Erlangen-Nuremberg, Erlangen, Germany.

I am grateful to Dora Söderberg for her support with various administrative matters. Special thanks to Karin Demin for her help, especially in the beginning of my stay here, when my knowledge of Swedish was nil.

My professors from India, Prof. C. Jaganmohan Rao, Prof. G. V. Prabhakar Rao and Prof. K. Tiwari, and Raghunath Sharma, deserve special mention for inspiring and encouraging me to walk the path of research.

My thanks to all my friends in India and Sweden, who enriched my life with their presence. I would not be what I am today without the support, care and love that I receive from my family. To Swami, Baba, Amma, Appa, Badri and Manni, words fail to express my gratitude.

Sriram Srinivasan
Stockholm, October 2005.

Contents

Abbreviations

ANC	Adaptive noise canceller
AR	Autoregressive
BM	Blocking matrix
BSS	Blind source separation
DCT	Discrete cosine transform
DFT	Discrete Fourier transform
DSB	Delay-and-sum beamformer
EM	Expectation maximization
EVRC	Enhanced variable rate coder
FBF	Fixed beamformer
FSB	Filter-and-sum beamformer
GLA	Generalized Lloyd Algorithm
GMM	Gaussian mixture model
GSC	Generalized sidelobe canceller
HMM	Hidden Markov model
KLT	Karhunen Loeve transform
LCMV	Linearly constrained minimum variance
LP	Linear prediction
LSA	Log-spectral amplitude
LSF	Line spectral frequency
MAP	Maximum a-posteriori

ML	Maximum-likelihood
MMSE	Minimum mean squared error
MSE	Mean squared error
MVDR	Minimum variance distortionless response
MWF	Multi-channel Wiener filter
pdf	Probability density function
PESQ	Perceptual evaluation of speech quality
PSD	Power spectral density
SD	Spectral distortion
SDB	Superdirective beamformer
SMV	Selectable mode vocoder
SNR	Signal-to-noise ratio
SSNR	Segmental signal-to-noise ratio
STFT	Short-time Fourier transform
STP	Short-term predictor
STSA	Short-time spectral amplitude
SVD	Singular value decomposition
VAD	Voice activity detector
VQ	Vector quantization

Part I

An Introduction to Speech Enhancement

Speech enhancement

1 Introduction

Speech is a fundamental means of human communication. Beginning with limited distance fixed-line telephone networks, recent developments have now given way to high quality mobile communication across the globe. Cellular phones and professional mobile radio such as those used by emergency services are an integral part of everyday life.

While the freedom and flexibility provided by mobile technology has made it possible to communicate outside controlled environments, it has also introduced new challenges. Mobile users communicate in different environments with varying levels and types of background noise such as traffic noise, car engine noise, multi-talker babble noise as in cafeterias etc. Suppression of the acoustic background noise is a relevant and challenging problem. Apart from reducing listener fatigue and improving the quality and intelligibility of speech, noise reduction is also crucial to obtain good performance of the speech coding algorithms that make mobile communication feasible.

Robustness to environmental noise has remained a limiting factor in the widespread deployment of speech enabled services such as speech recognition and speaker identification systems. While these technologies show impressive performance in controlled noise-free environments, performance rapidly degrades under practical noisy conditions. Noise reduction is also becoming an increasingly important feature in hearing aids. For these reasons, much effort has been devoted over the last few decades towards developing efficient speech enhancement algorithms. The term speech enhancement has a broad connotation encompassing various topics such as acoustic background noise reduction, dereverberation, blind source separation of speech signals, bandwidth extension of narrowband speech, etc. In this thesis, we use speech enhancement to describe acoustic background noise reduction.

Noise reduction can be viewed as an estimation problem, where an unknown signal (speech) is to be estimated in the presence of noise, where only the noisy observation is available. The estimation-theoretic view is also meaningful in the estimation of parameters derived from the clean speech signal such as the linear predictive (LP) coefficients. The first step towards obtaining a rigorous solution to any estimation problem is to define a mathematical model for the observed data. Often, to account for the random nature of the data, this is done by ascribing a probability density function (pdf) to the observed data. The pdf is parameterized by the unknown parameters to be estimated.

Prior knowledge about the desired signal and the background noise is encapsulated by the respective pdfs. A simple model to exploit prior knowledge is to ascribe a particular form to the pdf, which is arrived at based on a large data set. For example, the speech pdf may be described using a Laplacian density and the noise pdf using a Gaussian density. A more accurate method, though computationally more demanding, is to use more sophisticated statistical models using, e.g., hidden Markov models (HMMs), Gaussian mixture models (GMMs), or codebooks that have been trained using a representative database. The pdfs of the speech and noise processes are thus estimated from corresponding training sequences. This is the approach adopted in this thesis, where prior knowledge about the speech and noise signals, in the form of trained codebooks of their LP coefficients, is used in the estimation procedure.

Given a pdf for the observed data, it is possible to adopt one of two different schools of estimation depending on the assumptions on the unknown parameter. If the parameter is assumed to be deterministic (but unknown), the procedure is termed *classical* estimation, e.g., maximum-likelihood (ML) estimation. If we assume that the unknown parameter is a random variable with its own pdf, and we estimate a realization of that random variable, the procedure is termed *Bayesian* estimation. In the work performed in this thesis, both maximum-likelihood and Bayesian approaches are considered. A brief description of the two estimation approaches is thus in order and is provided in section 2.

The vast family of speech enhancement algorithms may be broadly classified into two categories: single and multi-channel enhancement. Single-channel methods operate on the input obtained from only one microphone. They have been attractive due to cost and size factors, especially in mobile communications. In contrast, multi-channel methods employ an array of two or more microphones to record the noisy signal and exploit the resulting spatial diversity. The two approaches are not necessarily independent, and can be combined to improve performance. For example, in practical diffuse noise environments, the multi-channel enhancement schemes rely on a single-channel post-filter to provide additional noise reduction.

We discuss single-channel methods and introduce the contributions of this thesis towards this area in section 3. This section is intended to be a survey on single-channel enhancement algorithms. Multi-channel approaches and aspects of the thesis in this context are discussed in section 4.

2 Estimation-theoretic approach

In this section, we briefly outline the principles behind maximum-likelihood and Bayesian MMSE estimation. Their use in the speech enhancement application is discussed. Besides providing an overview of these techniques, this section also establishes some of the notation used in the remainder of this thesis.

2.1 Maximum-likelihood estimation

Consider the estimation of a parameter $\theta = [\theta_1 \dots \theta_p]^T$ based on a sequence of K observations $\mathbf{y} = [y(0) \dots y(K-1)]^T$. In ML estimation, θ is treated as a deterministic variable. The ML estimate of θ is the value θ^{ML} that maximizes the likelihood function $p(\mathbf{y};\theta)$ defined on the data. ML estimation has several favorable properties, in particular, it is asymptotically unbiased and efficient, i.e., as the number of observations K tends to infinity, the ML estimate is unbiased and achieves the Cramer-Rao lower bound (CRLB). It can be shown (assuming that the derivatives of the log-likelihood exist) that [121, ch. 7]

$$\theta^{\mathrm{ML}} \underset{K\to\infty}{\sim} \mathcal{N}(\theta, \mathbf{I}^{-1}(\theta)), \tag{1}$$

where $\mathbf{I}(\theta)$ is the $p \times p$ Fisher information matrix whose $(i,j)^{th}$ entry is given by

$$[\mathbf{I}(\theta)]_{ij} = -\mathrm{E}\left[\frac{\partial^2 \ln p(\mathbf{y};\theta)}{\partial\theta_i \partial\theta_j}\right]. \tag{2}$$

Thus we have (asymptotically)

$$\begin{aligned} \text{Unbiased: } & \mathrm{E}[\theta^{\mathrm{ML}}] = \theta, \\ \text{CRLB: } & \mathrm{var}(\theta_i^{\mathrm{ML}}) = [\mathbf{I}^{-1}(\theta)]_{ii}. \end{aligned} \tag{3}$$

The maximization of the likelihood function is performed over the domain of θ. In many cases, θ^{ML} cannot be computed in closed form and a numerical solution is obtained instead. Such numerical solutions are typically obtained through iterative maximization procedures such as the Newton-Raphson method or the expectation-maximization (EM) approach. The initial value of the parameter used to start the iterative procedure usually has a strong impact on whether the final estimate results in a local or a global maximum of the likelihood function.

In applications where the parameter θ is known to assume one of a finite set of values, the problems due to the iterative procedures can be avoided by performing the maximization over this finite set. An exhaustive search over the finite parameter space guarantees a global maximum.

For speech enhancement, we assume that both speech and noise can be described by independent auto-regressive (AR) processes. The problem is then one of estimating the speech and noise LP coefficients[1] [6, 138] based on the observed noisy speech in an ML framework. The clean speech AR model can be mathematically expressed as

$$x(n) = \sum_{l=1}^{p} a_l x(n-l) + e(n), \tag{4}$$

where $a_1, \ldots, a_p$ are the LP coefficients of order p and $e(n)$ is the prediction error, also referred to as the excitation signal. It is common to model $e(n)$ as a Gaussian random process. The LP analysis is typically performed for each frame of 20-30 ms, within which speech can be assumed to be stationary. For each frame, the model parameters are the vector of LP coefficients $\theta = [a_1 \ \ldots \ a_p]$, and the variance of the excitation signal. A similar model can be obtained for the noise signal.

The physiology of speech production imposes a constraint on the possible shapes of the speech spectral envelope. Since the spectral envelope is specified by the LP coefficients [138], this knowledge can be modelled using a sufficiently large codebook of speech LP coefficients obtained from long sequences of training data. Such a-priori information about the LP coefficients of speech has been exploited successfully in speech coding using trained codebooks [174]. Similarly, noise LP coefficients can also be modelled based on training sequences for different noise types. Thus, it is sufficient to perform the maximization over the speech and noise codebooks. The search results in a global optimum in the constrained search space.

In paper A, we describe an ML approach for the estimation of the speech and noise LP coefficients and the excitation variances. Together, they characterize the speech and noise power spectra, which can be used to construct a Wiener filter to obtain the enhanced speech signal. Given the noisy data, the excitation variances maximizing the likelihood are determined for each pair of speech and noise LP coefficients from the codebooks. This is done for all combinations of codebook pairs, and the most likely codebook combination, together with the optimal excitation variances, is obtained. Since this optimization is performed on a frame-by-frame basis, good performance is achieved in nonstationary noise environments.

Apart from restricting the search space, using a codebook in the ML estimation has an additional benefit in applications where a codebook index needs to be transmitted over a network, e.g., in speech coding. In this case,

[1] LP and AR modelling are closely related. In LP, the goal is to determine a FIR filter that can predict a future sample as a linear combination of past samples in an optimal (squared error sense) fashion. The difference between the original and predicted signals is termed the prediction error, which for an AR signal is white noise. In AR modelling, the goal is to obtain an all-pole IIR filter, which when excited with white noise results in a signal whose statistics are the same as that of the signal being modelled. The variance of the prediction error in LP equals the variance of the excitation signal in AR modelling. In this thesis, we use the terms AR and LP parameters interchangeably.

the likelihood function can be interpreted as a modified distortion criterion to select the best codebook entry in the presence of noise. This approach is adopted using a multi-stage codebook in [207].

2.2 Bayesian MMSE estimation

In ML estimation, the parameter θ is treated as a deterministic but unknown constant. In the Bayesian approach, θ is treated as a random variable. The Bayesian methodology allows us to incorporate prior (before observing the data) knowledge about the parameter by assigning a prior pdf to θ. A cost function is formulated and its expected value, referred to as the Bayesian risk, is minimized. A commonly used cost function is the mean squared error (MSE). In this case, the Bayesian minimum mean squared error (MMSE) estimate θ^{BY} of θ given the observations $\mathbf{y}$ is obtained by minimizing $\mathrm{E}[(\theta - \theta^{\mathrm{BY}})^2]$, where E is the statistical expectation operator. The expectation is with respect to the joint distribution $p(\mathbf{y}, \theta)$. Thus, the cost function to be minimized can be written as

$$\begin{aligned}\eta &= \mathrm{E}[(\theta - \theta^{\mathrm{BY}})^2] \\ &= \int\int (\theta - \theta^{\mathrm{BY}})^2 p(\mathbf{y}, \theta) d\mathbf{y} d\theta \\ &= \int \left(\int (\theta - \theta^{\mathrm{BY}})^2 p(\theta|\mathbf{y}) d\theta \right) p(\mathbf{y}) d\mathbf{y}, \end{aligned} \tag{5}$$

where the posterior pdf $p(\theta|\mathbf{y})$ is the pdf of θ after the observation of data. Since $p(\mathbf{y}) \geq 0$, it is sufficient to minimize the inner integral for each $\mathbf{y}$. An estimate of θ can be found by determining a stationary point of the cost function (setting the derivative of the inner integral to zero). We can write

$$\frac{\partial}{\partial \theta^{\mathrm{BY}}} \int (\theta - \theta^{\mathrm{BY}})^2 p(\theta|\mathbf{y}) d\theta = 0 \tag{6}$$

so that

$$\theta^{\mathrm{BY}} = \int \theta p(\theta|\mathbf{y}) d\theta = \mathrm{E}[\theta|\mathbf{y}]. \tag{7}$$

For simplicity, in (7) and in the remainder of this thesis, we use the notation $\mathrm{E}[\theta|\mathbf{y}]$ instead of the more rigorous notation $\mathrm{E}[\theta|\mathbf{Y} = \mathbf{y}]$, where $\mathbf{y}$ is a realization of the corresponding random variable $\mathbf{Y}$. Using Bayes' rule, the posterior pdf can be written as

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}, \tag{8}$$

where the denominator $p(\mathbf{y})$ is a normalizing factor, independent of the parameter θ.

In paper B, we describe a method to obtain Bayesian MMSE estimates of the speech and noise AR parameters. The respective prior pdfs are modelled by codebooks. The integral in (7) is replaced by a summation over the codebook entries. We also consider MMSE estimation of functions of the AR parameters, and one such function is shown to result in the MMSE estimate of the clean speech signal, given the noisy speech. As in the ML case, MMSE estimates of the speech and noise AR parameters are obtained on a frame-by-frame basis, ensuring good performance in nonstationary noise.

In the ML estimation framework, one pair of speech and noise codebook vectors was selected as the ML estimate, whereas the Bayesian approach results in a weighted sum of the speech (noise) codebook vectors. The Bayesian method provides a framework to account for both the knowledge provided by the observed data and the prior knowledge.

3 Single-channel speech enhancement

Single-channel speech enhancement systems obtain the input signal using only one microphone. This is in contrast to multi-channel systems where the presence of two or more microphones enables both spatial and temporal processing. Single-channel approaches are relevant due to cost and size factors. They achieve noise reduction by exploiting the spectral diversity between the speech and noise signals. Since the frequency spectra of speech and noise often overlap, single-channel methods generally achieve noise reduction at the expense of speech distortion.

The reduction of background noise using single-channel methods requires an estimate of the noise statistics. Early approaches were based on voice activity detectors (VAD), e.g., [74], and noise estimates were updated during periods of speech inactivity. Accuracy deteriorates with decreasing signal-to-noise ratios (SNR) and in nonstationary noise. Soft-decision VADs [139, 183, 194, 195] update the noise statistics even during speech activity. A number of other noise estimation methods have been proposed [39, 43, 44, 54, 55, 91, 144, 184, 210, 224, 231] and are discussed in section 3.1.

Since single-channel methods exploit the spectral diversity between the speech and noise signals, it is therefore natural to perform the processing in the frequency domain. Processing is done on short segments of the speech signal, typically of the order of 20 to 30 ms, to ensure that the speech signal satisfies assumptions of wide-sense stationarity. The segmentation is performed using a sliding window of finite support. The windowed signal (assuming it is absolute summable) is transformed to the frequency domain using the discrete short-time Fourier transform (STFT) [180, ch. 7]:

$$X_m(k) = \frac{1}{\sqrt{K}} \sum_{n=-\infty}^{\infty} x(n)h(n-m)\exp(-j\frac{2\pi}{K}kn), \;\; k = 0, 1, \ldots, K-1, \quad (9)$$

where $x(n)$ is the sampled speech signal, $h(n)$ is the analysis window that is non-zero only in the interval $[0, K-1]$, m is the index to the current windowed segment, and k is the discrete frequency index[2]. While it is not customary to include the normalization by $\sqrt{K}$ in the definition, we do so for convenience in notation introduced later in the thesis. To obtain $X_{m+1}(k)$, the window is shifted by one sample from its previous position. In practice, the sequence of frames is subsampled by a factor L, resulting in $X_{mL}(k)$, which is equivalent to a larger frame-shift. Typical values at a sampling frequency of 8 kHz are $K = 256$ (32 ms) and a frame-shift of $L = 128$ (50% overlap). For a given window length K over which the analysis window is non-zero, to ensure invertibility of the discrete STFT, we must have $L \leq K$. In practice, (9) is implemented by buffering K samples of the signal, applying a smooth window, followed by a K-point discrete (fast) Fourier transform (DFT). For the next frame, the buffer is advanced by L samples.

We now introduce some notation and terminology. In the remainder of this chapter, we drop the frame index m, and the processing is described for a single frame. We refer to $X(k)$ as the (complex) spectrum of the signal and to $|X(k)|$ as the magnitude spectrum. The quantity $|X(k)|^2$ denotes the periodogram. For stationary signals, as $K \rightarrow \infty$, the expected value of the periodogram can be shown to be the power spectral density (PSD), $P_x(k) = \mathrm{E}\{|X(k)|^2\}$. The PSD and the autocorrelation function of the signal form a Fourier transform pair.

We consider an additive noise model

$$y(n) = x(n) + w(n), \tag{10}$$

where $y(n)$ represents the sampled noisy speech. The speech and noise signals are modelled as independent random processes. Let $\mathbf{x} = [x(0)\, x(1) \ldots x(K-1)]^T$ denote a segment of length K of the clean speech signal. $\mathbf{y}$ and $\mathbf{w}$ are defined analogously. In the noise reduction problem, we wish to obtain an estimate $\hat{\mathbf{x}}$ of the clean speech from the noisy observation. The additive signal model defined above applies to all the single-channel algorithms described in this thesis.

The additive model can be expressed in the frequency domain as

$$Y(k) = X(k) + W(k), \tag{11}$$

where $Y(k), X(k)$ and $W(k)$ are obtained by applying the DFT to the time-domain entities $\mathbf{y}, \mathbf{x}$ and $\mathbf{w}$ respectively. Since the speech and noise signals

[2] The discrete STFT is obtained by sampling the STFT, which is continuous in frequency, with a frequency sampling interval of $\frac{2\pi}{N}$, i.e., $X_m(k) = X_m(\omega)_{|\omega=\frac{2\pi}{N}k}$, $k = 0, 1, \ldots, N-1$, where N is the frequency sampling factor. To ensure perfect reconstruction, we must have $N \geq K$. For simplicity, in this thesis, we assume $N = K$.

are independent, the following relation holds between the corresponding PSDs:

$$P_y(k) = P_x(k) + P_w(k). \tag{12}$$

A block diagram of a generic frequency domain single-channel speech enhancement system is shown in Fig. 1. It is common to modify only the spectral amplitude and use the noisy phase [218]. An estimate of the noise PSD is obtained from the noisy speech. Any available prior knowledge about the noise signal may be exploited. Using the noise estimate, an estimate of the spectral coefficients of clean speech is obtained from the noisy coefficients. Again, prior knowledge about the speech signal or about the human auditory system can be exploited. In some systems (e.g., the systems described in papers A and B), the speech and noise PSD are jointly estimated, as indicated by the bidirectional arrow in the figure. Enhanced speech is reconstructed in the time domain through the inverse discrete Fourier transform (IDFT) and through an overlap-add technique. We note that the overlap in the enhancement system results in an algorithmic delay. This delay can be minimized through careful combination with the analysis/synthesis schemes of the speech coders that the enhancement systems may be part of [149].

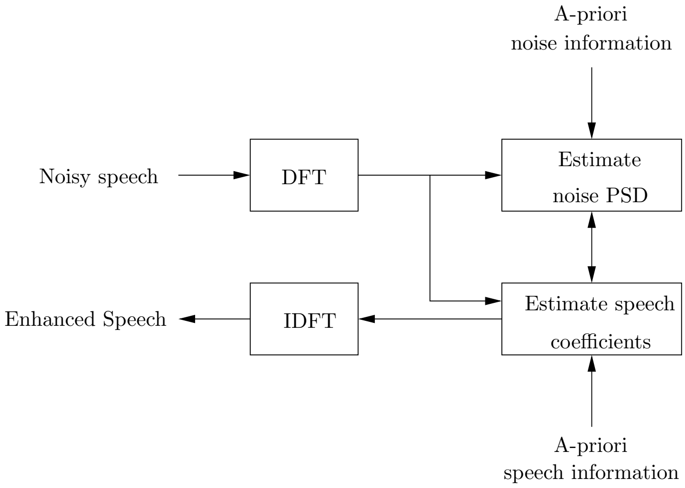

Figure 1: Block diagram of frequency domain single-channel speech enhancement.

Several different single-channel techniques have been developed over the last few decades, and we provide an overview in this chapter. Section 3.1 discusses methods to estimate the noise PSD. Wiener filtering, spectral subtraction, subspace based methods and Kalman filter methods, their similarities and differences, are discussed in section 3.2. Section 3.3 is devoted to approaches that assume a statistical distribution on particular representations of the speech signal. Systems that exploit a-priori information about speech and noise signals through the use of trained statistical models are considered in section 3.4. While the above sectioning provides an implicit categorization of speech enhancement techniques, we emphasize that this categorization is by no means comprehensive and is solely for convenience.

3.1 Noise estimation

Estimation of the statistics of the background noise is an essential feature of single-channel noise reduction algorithms. It is a challenging task as the estimates have to be obtained from the noisy speech signal. A common approach is to use a voice activity detector (VAD) to identify time segments in the signal where speech is absent and thus the signal consists of only the background noise [19, 20, 37]. Estimates of the noise statistics are updated during these speech pauses. While VAD based noise estimation schemes have the advantage of low computational complexity, they suffer from two problems. First, with decreasing signal-to-noise ratios, detecting speech pauses is no longer a trivial task. Second, while the method works reasonably well in stationary noise environments, performance degrades in environments where the noise statistics continuously change, which is often the case in practice.

To address the shortcomings of binary VADs (there are only two states, speech presence and absence), soft-decision VADs [35, 36, 139, 183, 194, 195] were proposed that assign a probability of speech presence to each segment. Thus, it is possible to update the noise statistics continuously, based on the probability that speech is present.

One of the noise estimation schemes that adapts also during speech activity is the minimum statistics approach [142, 144]. This method relies on the observation that the PSD of the noisy signal often decays to that of the noise signal. By maintaining a finite buffer of the smoothed noisy signal power spectra over time, noise estimates can be obtained by tracking the minimum in the buffer for each frequency bin. The smoothing factor is time-frequency dependent and optimally derived by minimizing an appropriate error criterion. Furthermore, since the minimum of a set of noisy PSD values is generally smaller than the mean, the method incorporates a bias compensation scheme to improve the estimation.

To reduce complexity, the buffer is decomposed into smaller buffers over which the minimum is tracked. To improve performance in nonstationary noise, the method contains provisions to detect local minima in the sub-windows in vicinity of the overall minimum. To ensure that the noisy PSD indeed decays to that of the noise signal for each frequency bin, the buffer has to be sufficiently large. However, large buffers make it difficult to adapt to quickly changing noise. Thus, there is a trade-off between accuracy of the estimates and adaptation to nonstationary noise. While the method provides good performance in stationary noise environments, performance in highly nonstationary noise environments is limited by the buffer size.

A related scheme is the quantile method described in [210]. Similar to the minimum statistics approach, this method is also based on order statistics, and maintains a buffer of past power spectra (in [210], no smoothing is performed, and the periodogram is stored in the buffer). However instead of considering the minimum, the noise estimate is obtained as the q^{th} quantile of the values in the buffer for each frequency bin ($q = 0$ corresponds to the minimum). This method suffers from the same limitations as the minimum statistics approach.

The minima controlled recursive averaging approach introduced in [39, 43, 44], also similar to the minimum statistics method [144], tracks the minima of the recursively averaged noisy power spectrum. In [39], two iterations of smoothing and minimum tracking are performed. The first iteration serves as a rough VAD which helps to eliminate strong speech components in the smoothing in the second iteration. Thus smaller buffers are sufficient. However, while the method has an advantage in nonstationary noise compared to [144], performance is still limited by the buffer size.

Recursive noise estimation algorithms for nonstationary noise environments have been proposed in the cepstral domain in the context of speech recognition. Employing a GMM to describe speech, the time-varying noise parameters (considered to be deterministic) are obtained using iterative stochastic approximation in [52, 54]. The recursive estimation employs a forgetting factor that introduces a tradeoff between the accuracy of the estimate and the speed with which changes are tracked. Following a similar recursive estimation framework, maximum a-posteriori (MAP) estimates of the noise parameters in the log-domain are obtained in [51] by employing a Gaussian prior for the noise. The mean and the covariance of the prior are fixed and obtained using initial noise-only segments. This approach was improved in [53] by including adaptation of the noise prior by recursively updating the prior statistics.

One way to overcome the limitations of noise estimation methods in nonstationary environments is to use data-driven prior information about noise, when available. Such an approach is suggested for example in [188], where HMMs are used to model the noise power spectra, characterized by the LP coefficients and the variance of the excitation (gain). Different HMMs are trained for different noise types. Based on the observation, an appropriate noise model is selected. Since the trained noise model includes the gain, such a method requires a gain adaptation scheme that adjusts the gain based on the observation, as conditions generally differ during training and testing. In [188], this gain adaptation is performed using an estimate of the noise gain obtained from silence segments. An obvious improvement is to use a more robust long-term gain estimate based on the minimum statistics approach, which also updates during speech activity. Such HMM based methods can handle changes to the noise spectral shape, which are well modelled by the noise HMMs. However, they can adapt only as quickly as the long-term estimate [144] to changes in the gain.

A method that computes the noise gain on a frame-by-frame basis can respond quickly to changes. Such a method for noise reduction is discussed in [125, 211]. In [125], the speech and noise power spectra are described by two trained codebooks of the respective AR coefficients. Unlike the HMM method where the excitation variance is part of the prior information, the codebooks do not contain the excitation variance, which is then computed for each input frame using the noisy observation at hand. This method was extended and presented in an ML framework in [205] with an optimal (in the ML sense) estimation of the excitation variances. In [205], multiple noise codebooks are used, and for each frame, one noise codebook is selected according to an appropriate criterion for subsequent use in the ML estimation. A codebook based Bayesian MMSE approach with instantaneous gain estimation is described in [209], where it is shown to result in better performance than the ML approach and the HMM based MMSE approach. The price to be paid for the improved performance in nonstationary environments is an increase in computational complexity compared to methods such as [144].

Methods that use prior knowledge, such as [125, 126, 135, 188, 205–209], are discussed in greater detail in section 3.4 in conjunction with the speech enhancement scheme that they are associated with.

3.2 The Wiener filter and its relatives

Wiener filtering, spectral subtraction, subspace methods and Kalman filtering are popularly used approaches for noise reduction. In the following subsections, we discuss these methods, and study their differences and similarities.

Wiener filtering

Consider a $K \times K$ linear estimator H that results in an estimate $\hat{\mathbf{x}} = H\mathbf{y}$ of the clean speech from the noisy speech. The estimation error can be written as

$$\boldsymbol{\epsilon} = H\mathbf{y} - \mathbf{x} = \underbrace{(H - I)\mathbf{x}}_{\boldsymbol{\epsilon}_{\mathbf{x}}} + \underbrace{H\mathbf{w}}_{\boldsymbol{\epsilon}_{\mathbf{w}}}, \tag{13}$$

where I is the $K \times K$ identity matrix. The mean squared estimation error is given by $\operatorname{tr} \mathrm{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\}$, where tr denotes the matrix trace. From the independence assumption, the cross terms in the mean squared error vanish and we have

$$\begin{aligned}\operatorname{tr} \mathrm{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\} &= \operatorname{tr} \mathrm{E}\{\boldsymbol{\epsilon}_{\mathbf{x}}\boldsymbol{\epsilon}_{\mathbf{x}}{}^{\mathrm{T}}\} + \operatorname{tr} \mathrm{E}\{\boldsymbol{\epsilon}_{\mathbf{w}}\boldsymbol{\epsilon}_{\mathbf{w}}{}^{\mathrm{T}}\} \\ &= \operatorname{tr} (H-I)R_x(H-I)^T + \operatorname{tr} HR_wH^T, \end{aligned} \tag{14}$$

where $R_x = \mathrm{E}\{\mathbf{x}\mathbf{x}^{\mathrm{T}}\}$ and $R_w = \mathrm{E}\{\mathbf{w}\mathbf{w}^{\mathrm{T}}\}$ are the covariance matrices of speech and noise respectively.

In the Wiener filter approach, the optimal estimator is designed to minimize the mean squared error, i.e.,

$$\begin{aligned} H_W &= \arg\min_H \left(\operatorname{tr} (H-I)R_x(H-I)^T + \operatorname{tr} HR_wH^T\right) \\ &= R_x(R_x+R_w)^{-1}. \end{aligned} \tag{15}$$

The above estimator can be efficiently implemented in the frequency domain. Under the assumption of large K, the covariance matrices R_x and R_w can be approximated as circulant and are hence diagonalized by the DFT, i.e, $R_x = F^*P_xF$, where $P_x = \mathrm{diag}(P_x(0)\ P_x(1)\dots P_x(K-1))$ is a diagonal matrix containing the PSD of $\mathbf{x}$, F is the DFT matrix, and the superscript * denotes complex conjugate transpose. Similarly, $R_w = F^*P_wF$, where $P_w = \mathrm{diag}(P_w(0)P_w(1)\dots P_w(K-1))$ is a diagonal matrix containing the PSD of $\mathbf{w}$. With the above diagonalization, the Wiener filter can be rewritten in the frequency domain as

$$H_W^{\mathrm{freq}} = P_x(P_x+P_w)^{-1} = P_xP_y^{-1}. \tag{16}$$

However, P_x is not known, and, in practice, an estimate $\hat{P}_x$ of P_x is used. This estimate is commonly obtained in a subtractive fashion from (12) using P_y and an estimate $\hat{P}_w$ of P_w, and negative values are set to zero since the PSD cannot be negative (negative values may arise since $\hat{P}_w$ is only an estimate of the noise PSD):

$$\hat{P}_x(k) = \max(P_y(k) - \hat{P}_w(k), 0) \quad k = 0, 1, \dots, K-1, \tag{17}$$

so that the clean speech spectrum is then estimated according to

$$\hat{X}(k) = \frac{\max(P_y(k) - \hat{P}_w(k), 0)}{P_y(k)} Y(k) \quad k = 0, 1, \dots, K-1. \tag{18}$$

In practice, $P_y(k)$ is approximated using the periodogram or a smoothed version there-of.

Spectral subtraction

Spectral subtraction is a speech enhancement scheme based on a direct estimation of the short-time spectral magnitude of clean speech [17]. The estimated magnitude is combined with the noisy phase. This can be written as

$$\hat{X}(k) = \max(|Y(k)| - \overline{|W(k)|}, 0)\frac{Y(k)}{|Y(k)|}, \tag{19}$$

where $|\hat{X}(k)|$ is estimated by subtracting $\overline{|W(k)|}$, an average estimate of the magnitude spectrum of the noise signal, from the noisy spectral magnitude. Negative values resulting from the subtraction are set to zero since the magnitude spectrum cannot be negative.

In one variant, called power spectral subtraction, an estimate of the periodogram of the clean speech signal is obtained as

$$|\hat{X}(k)|^2 = \max(|Y(k)|^2 - P_w(k), 0), \tag{20}$$

the square root of which is then combined with the noisy phase:

$$\hat{X}(k) = \sqrt{\max(|Y(k)|^2 - P_w(k), 0)}\frac{Y(k)}{|Y(k)|}. \tag{21}$$

The name power spectral subtraction is a misnomer since the periodogram of the noisy speech is used in the subtraction and not the PSD. The name arose due to the close similarity between (17) and (20) [132].

Power spectral subtraction is derived from an ML perspective in [156], where $Y(k)$ is assumed to be a complex zero mean Gaussian random variable with variance $P_y(k) = P_x(k) + P_w(k)$. Thus, its real and imaginary parts are both zero mean Gaussian with variance $P_y(k)/2$. The resulting pdf can be written as

$$p(Y(k)) = \frac{1}{\pi(P_x(k) + P_w(k))} \exp\left(-\frac{|Y(k)|^2}{P_x(k) + P_w(k)}\right) \tag{22}$$

so that the ML estimate of $P_x(k)$ is obtained as

$$\hat{P}_x(k) = |Y(k)|^2 - P_w(k). \tag{23}$$

Negative values are set to zero. The clean speech spectrum is then obtained according to [156]:

$$\hat{X}(k) = \sqrt{\hat{P}_x(k)}\frac{Y(k)}{|Y(k)|} = \sqrt{\frac{\max(|Y(k)|^2 - P_w(k), 0)}{|Y(k)|^2}}Y(k), \tag{24}$$

which is identical to (21). If the noisy PSD $P_y(k)$ is approximated by the periodogram $|Y(k)|^2$, it can be seen that the Wiener filter (18) corresponds to the square of the suppression rule for power spectral subtraction.

One of the main drawbacks of the spectral subtraction scheme is that the enhanced signal suffers from *musical* noise, which is especially audible in speech pauses. Random fluctuations in the periodogram result in randomly spaced peaks in the enhanced spectrum, after the spectral attenuation. In between these peaks, the spectral values are strongly attenuated since they are close to or below the estimated noise PSD. In the time domain, this residual noise is perceived as a sum of pure tones corresponding to the peaks, and is hence referred to as musical noise. The musical noise phenomenon is common to many frequency domain speech enhancement algorithms, e.g., the Wiener filter also suffers from this problem since, in practice, $\hat{P}_x(k)$ is often obtained in a subtractive fashion using (17). Magnitude estimation schemes that avoid musical noise [28, 68, 69] are discussed in section 3.3.

The error arising from the use of the noisy periodogram $|Y(k)|^2$ instead of the PSD $P_y(k)$ in (23) is analyzed in detail in [97]. While the periodogram is an asymptotically unbiased estimate of the PSD, its variance does not tend to zero even as the frame length K approaches infinity. For stationary noise, the variance can be reduced by estimating the PSD as an average of the periodograms from multiple frames. For speech, the variance can be reduced by employing a parametric estimate, e.g., by using an AR model as in (4) so that the variance is proportional to $\frac{2p}{K}$, where p is the AR model order (typically, $p = 10$ for narrowband speech) [97].

Several modifications have been made to the basic spectral subtraction scheme. Often a generalized subtraction rule of the form [89, 132, 134, 191, 217]

$$\hat{X}(k) = \left(1 - \beta_k \frac{[P_w(k)]^{\alpha/2}}{|Y(k)|^{\alpha}}\right)^{\frac{1}{\alpha}} Y(k), \tag{25}$$

is employed, with provisions to prevent the occurrence of negative amplitude values due to the subtraction. The amount of subtraction (and hence the musical noise) is controlled by β_k. $\alpha = 1$ corresponds to amplitude spectral subtraction and $\alpha = 2$ to power spectral subtraction. A non-linear choice for β_k (viewed as a function) is suggested in [134], with $\alpha = 1$. In [217], the parameter values are determined using the auditory masking threshold.

Subspace based methods

In equation (13), the estimation error $\boldsymbol{\epsilon}$ is the sum of two components, $\boldsymbol{\epsilon}_{\mathbf{x}}$, which represents the distortion introduced into the speech signal, and $\boldsymbol{\epsilon}_{\mathbf{w}}$, which is the residual noise remaining after the enhancement. The Wiener filter does not make a distinction between the two types of distortions. An alternate approach motivated by perceptual considerations is to have a trade-off between noise reduction and signal distortion, and was introduced in connection with the so-called subspace methods. In [70] for instance, the goal is to minimize the speech distortion $\operatorname{tr} \mathrm{E}\{\boldsymbol{\epsilon}_{\mathbf{x}}\boldsymbol{\epsilon}_{\mathbf{x}}^{\mathrm{T}}\}$ subject to a constraint[3] on the residual noise level tr $\mathrm{E}\{\boldsymbol{\epsilon}_{\mathbf{w}}\boldsymbol{\epsilon}_{\mathbf{w}}{}^{\mathrm{T}}\}$. This leads to a constrained optimization problem that can be solved by the Lagrange multiplier method. The resulting estimate can be written as

$$H_S = R_x(R_x + \mu R_w)^{-1}, \tag{26}$$

where $\mu > 0$ is the Lagrange multiplier. As observed in [70], for $\mu = 1$, H_S coincides with the Wiener filter H_W.

The subspace methods differ from the Wiener filter in the use of the data-dependent Karhunen Loeve transform (KLT) to diagonalize the covariance matrix, which avoids the approximation when the data-independent DFT is used. Let $R_y = U\Lambda_y U^T$ be the eigendecomposition of R_y, where U is an orthonormal matrix of the eigenvectors of R_y and Λ_y is a diagonal matrix containing the corresponding eigenvalues. For white noise, R_w is diagonal, and constant across the diagonal. Thus, U^T, which is the KLT and therefore diagonalizes R_y, also diagonalizes R_x. Thus, we have

$$H_S = U\Lambda_x(\Lambda_x + \mu\Lambda_w)^{-1}U^T, \tag{27}$$

where Λ_x and Λ_w are diagonal matrices containing the eigenvalues of R_x and R_w respectively. We assume that the eigenvalues are sorted in nonincreasing order. The subspace approach also exploits the fact that the noisy signal space can be decomposed into a signal-plus-noise subspace that contains both the clean signal and noise, and a noise subspace that contains only noise (this implies the assumption that $\mathrm{rank}(R_x) = \mathrm{M} \leq \mathrm{rank}(R_y)$). Components of the noise subspace are nulled and components of the signal-plus-noise subspace are modified by a gain function of the form (27). The resulting estimator can be written as

$$H_{\mathrm{Sub}} = U \begin{bmatrix} G_\mu & 0 \\ 0 & 0 \end{bmatrix} U^T, \tag{28}$$

where $G_\mu = \Lambda'_x(\Lambda'_x + \mu\sigma_w^2 I)^{-1}$, Λ'_x is an $M \times M$ diagonal matrix containing only the first M eigenvalues of R_x, σ_w^2 is the variance of the white noise, and I denotes the $M \times M$ identity matrix. Note that the components of the noise subspace are nulled. Determination of the model subspace order M is addressed in [70].

The colored noise problem has been handled in different ways by different researchers. A pre-whitening transformation is suggested in [70]. In [165], a distinction is made between speech dominated frames and noise dominated frames. In speech dominated frames, the KLT matrix of R_x is used, and in noise dominated frames, the KLT matrix of R_w is used. This approach was shown to be better than the pre-whitening suggested in [70].

[3] The subspace methods discuss constraints both in the time domain and the spectral domain. We restrict our discussion to the time domain.

In [185], a diagonality assumption is imposed on the covariance matrix of the noise vectors in the KLT domain. A joint diagonalization of R_x and R_w is proposed in [107]. The matrix that achieves this joint diagonalization is the eigenvector matrix of $R_w^{-1} R_x$. This approach avoids the suboptimality due to the diagonal approximation in [185]. In [130], the whitening approach suggested in [70] is elaborated. The matrix H_S is simplified as $H_S = R_w^{1/2} U^* \Lambda(\Lambda + \mu I) U R_w^{-1/2}$, where U and Λ are the matrices of eigenvectors and eigenvalues of the whitened matrix $R_w^{-1/2} R_x R_w^{1/2}$. As noted in [130], since $R_w^{-1} R_x$ and $R_w^{-1/2} R_x R_w^{1/2}$ are similar matrices, they have the same eigenvalues and the Wiener-type gain modification performed in [107] (joint diagonalization of R_x and R_w) is identical to the whitening approach suggested in [70, 130].

Instead of applying the KLT to the noisy covariance matrix, a singular-value decomposition (SVD) can be applied to a Hankel (or Toeplitz) matrix formed from the noisy signal [50]. From a numerical point of view, this approach has the advantage that it does not require the computation of the covariance matrix. Singular values smaller than a threshold are set to zero to obtain a matrix with reduced rank. The output signal is reconstructed from the resulting matrix (after ensuring a Hankel structure through averaging). Pre-whitening of the signal is used to handle colored noise. The pre-whitening can also be included as an integral part of the algorithm using the quotient SVD of the matrix pair corresponding to the noisy and noise matrices. This approach is adopted for general broadband noise in [114]. The case of narrowband noise is addressed in [115].

Various modifications have been proposed to the basic subspace approach. Methods that incorporate psychoacoustic properties such as auditory masking into subspace based speech enhancement systems are described in [106, 112, 122, 123, 216]. The basic idea in these methods is to solve a constrained optimization problem that minimizes signal distortion while constraining the residual noise to lie below the masking threshold. The subspace method has also been successfully used as a front-end for speech recognition systems operating in noisy environments [103, 108–110].

Kalman filtering

Wiener filtering, spectral subtraction and the subspace methods discussed above can generally be categorized as non-parametric methods in the sense that they do not employ any parametric model to describe the speech signal. This is in contrast with parametric methods that use models such as the AR or the sinusoidal model to describe the signal. We discuss here one specific approach, the Kalman filter, which provides a framework that can exploit information about the human speech production process by using the AR model[4]. Since the first use of the Kalman filter approach for speech enhancement in [173], several researchers have proposed different techniques [76–79,85,90,92,129,178,197,226]. We briefly outline the general idea behind these approaches. We state the expression for the p^{th} order AR model again for convenience:

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + v(n), \tag{29}$$

where $a_1, \ldots, a_p$ are the LP coefficients and $v(n)$ is a zero mean white Gaussian process with variance σ_v^2. The additive noise model can then be expressed in a state-space form as

$$\begin{aligned} \mathbf{x}(n) &= \mathbf{F}\mathbf{x}(n-1) + \mathbf{g}\mathbf{v}(n) \\ y(n) &= \mathbf{h}^T\mathbf{x}(n) + w(n), \end{aligned} \tag{30}$$

where

$$\begin{aligned} \mathbf{x}(n) &= [x(n-p+1)\ x(n-p+2)\ldots x(n)]^T, \\ \mathbf{g} = \mathbf{h} &= [0\ 0\ \cdots 1]^T, \\ \mathbf{F} &= \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ . & . & . & \ldots & . & . \\ . & . & . & \ldots & . & . \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_2 & a_1 \end{bmatrix}, \end{aligned} \tag{31}$$

and $\mathbf{v}(n)$ is defined analogously to $\mathbf{x}(n)$. The linear state-space representation (30) provides a natural framework to incorporate the AR model. The case of white background noise was considered in [173] and the extension to colored noise in [85] where an AR-model is assumed for the noise as well.

Given the above state-space representation, the standard Kalman filter update equations can be employed to obtain an estimate of the clean speech [85]. It is shown in [92, 173] that the fixed-lag Kalman smoother with a lag well within the tolerable delay results in better performance than the causal Kalman filter. The performance of the causal Kalman filter can also be improved through pre- and post-processing to achieve a perceptual shaping of the residual noise [92].

Discussion

The Wiener filter, spectral subtraction, subspace methods and the Kalman filter exhibit strong similarities. One example is the concept of the trade-off

[4] We note that it is also possible to use the AR model representation of speech in the Wiener filter by specifying the speech PSD P_x in (16) with the AR parameters, see e.g., [131].

between the signal distortion and residual noise, which was introduced explicitly first for the subspace method [70]. This trade-off has been employed in spectral subtraction schemes in various forms such as under-subtraction [127] and over-subtraction [10], albeit in an ad-hoc manner. These ad-hoc fixes to improve performance were given a rigorous mathematical background by the class of subspace methods discussed in [70, 107, 165, 185].

The subspace methods decompose the noisy signal space into signal-plus-noise subspace and a noise subspace. Components of the noise subspace are nulled. A similar operation is performed in both Wiener filtering and spectral subtraction as seen in (17), (19) and (20), where subtracting an estimate of the noise PSD results in certain components being set to zero. This can be seen as nulling the components of the noise subspace. While the subspace methods use the KLT, the DFT is used in the other two methods.

Other transforms such as wavelets and wavelet packets have also been used in noise reduction. In contrast to the discrete STFT that uses a fixed analysis window, the wavelet transform uses short windows at high frequencies and long windows at low frequencies [186]. This leads to a better time resolution at high frequencies, which can be useful, e.g., in preserving transients [190], and is also closely related to human perception. The principle under which the wavelet-based methods operate is also similar to the subspace concept. The wavelet based methods achieve noise reduction through thresholding, which relies on the fact that only a few wavelet coefficients correspond to the signal. There are two types of thresholding - hard and soft. The hard thresholding operation on the noisy wavelet coefficient $\mathcal{W}_y$ is defined as [60]

$$\eta_H(\mathcal{W}_y, \lambda) = \begin{cases} \mathcal{W}_y & \text{if } |\mathcal{W}_y| > \lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{32}$$

This operation is similar to setting the components of the noise subspace to zero. The optimal (MMSE sense) threshold for white Gaussian noise can be derived as $\lambda = \sigma\sqrt{2\log K}$, where σ is the standard deviation of the noise and K is the number of samples in the observation [60]. For colored noise, a level dependent threshold was obtained in [117] as $\lambda_j = s_j\sqrt{2\log TK}$ where s_j is the standard deviation of the wavelet coefficients at level j of the transform.

The soft thresholding operation is defined as [59]

$$\eta_S(\mathcal{W}_y, \lambda) = \operatorname{sgn}(\mathcal{W}_y)\max(|\mathcal{W}_y| - \lambda, 0), \tag{33}$$

which can be viewed as setting the components of the noise subspace to zero, and performing a magnitude subtraction in the speech-plus-noise subspace.

Speech enhancement through thresholding in a wavelet packet domain is addressed in [16, 220]. While the basis (and hence the time-frequency tiling) is fixed in a wavelet transform, a wavelet packet transform selects the *best* basis from a library of orthonormal bases, by optimizing a certain cost function. The cost function is typically designed to concentrate the signal energy in a small number of transform coefficients [45]. Such an approach is applied to speech enhancement in [16] by subjecting the wavelet packet coefficients of the noisy signal to a hard thresholding operation. A modified hard thresholding based on the μ-law logarithm is proposed in [32]. A combination of soft and μ-law thresholding is proposed in [128].

3.3 Statistical model based systems

In the previous section, we considered linear estimation techniques for the signal. Linear estimation is optimal (in the MSE sense) for the case when $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian [118, ch. 3]. The Wiener filter represents the optimal solution in this case. In this section, we look at methods that use distributions other than Gaussian and derive optimal nonlinear solutions.

We first consider methods that retain the Gaussian assumption on the speech and noise processes in the frequency domain, i.e., the respective DFT coefficients are assumed to be normally distributed. They differ from the Wiener solution in that they attempt to obtain MMSE estimates of the spectral amplitude, which then follows a Rayleigh distribution. Next, we discuss methods that assume super-Gaussian (Gamma, Laplace etc.) models.

Gaussian models

In the Wiener filter approach to speech enhancement, an optimal (under the Gaussian assumption and in the mean squared error sense) estimate of the clean speech spectral component is obtained from the noisy speech. In [68], it is argued that the spectral amplitude is perceptually more relevant than the phase and thus performance could be improved by an optimal estimate of the amplitude. The amplitude estimate provided by the Wiener filter (obtained as the modulus of the optimally estimated spectral component) is not optimal under the assumed model; only the estimate of the spectral component is optimal. Using the same statistical model, an optimal estimate of the spectral amplitude, given the noisy speech, is obtained in [68]. The Fourier expansion coefficients of the speech and noise processes are assumed to be independent zero mean Gaussian variables with time-varying variances. This results in a Rayleigh distribution for the amplitudes of the Fourier coefficients.

The MMSE estimate of the complex exponential of the phase is also derived in [68] so that it can be used together with the MMSE amplitude estimate. It is shown that the modulus of the resulting estimate of the phase is not unity. Thus combining the MMSE phase estimate with the MMSE amplitude estimate affects the optimality of the amplitude. To address this problem, a constrained MMSE estimate of the phase is obtained, whose modulus is constrained to be unity. The resulting constrained MMSE estimate is the noisy phase itself. The MMSE amplitude estimate is obtained by applying the following gain function to the noisy spectral magnitude:

$$H_{\mathrm{EM}}(k) = \frac{\sqrt{\pi v_k}}{2\gamma_k} \exp\left(\frac{-v_k}{2}\right) \left[(1+v_k) I_0(\frac{v_k}{2}) + v_k I_1(\frac{v_k}{2})\right], \tag{34}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ are the modified Bessel functions of order zero and one respectively and

$$\begin{aligned} v_k &= \frac{\xi_k}{1+\xi_k}\gamma_k, \\ \xi_k &= \frac{P_x(k)}{P_w(k)}, \\ \gamma_k &= \frac{|Y(k)|^2}{P_w(k)}. \end{aligned} \tag{35}$$

The terms ξ_k and γ_k are referred to as the a-priori and a-posteriori SNR respectively [68, 156]. The MMSE spectral amplitude estimator [68] was shown to perform better than an ML estimate [156] using the same statistical model.

The concept of soft-decision noise suppression [156, 163] may be employed together with the MMSE amplitude estimate (34). In [68], a two-state model of speech presence/absence is used. The speech signal is assumed to be present in each spectral component with a probability $p = 0.5$.

In [196], using a statistical model similar to [68], and using the uncertainty of speech presence, an MMSE amplitude estimator is developed in the discrete cosine transform (DCT) domain. It has been shown through experiments that the DCT provides better energy compaction than the DFT [196, 225]. While the KLT provides optimal energy compaction[5], the DCT is computationally less demanding.

The noise suppression rule (34) proposed in [68] requires the computation of Bessel functions. Using the same statistical model, three simpler noise suppression rules that exhibit a behavior similar to (34) are derived in [223]. The three rules correspond to joint maximum a-posteriori (MAP) estimation of the amplitude and phase, MAP estimation of the amplitude, and MMSE estimation of the spectral power. The spectral power estimator was found to provide the best approximation to (34) and the corresponding estimator of the clean speech spectral component is computationally simpler:

$$H_{\mathrm{SP}}(k) = \sqrt{\frac{\xi_k}{1+\xi_k}\frac{1+v_k}{\gamma_k}}. \tag{36}$$

[5] Asymptotically, as the frame length $K \to \infty$, the energy compaction property of the DFT and the DCT approaches that of the KLT.

Motivated by the observation that the MSE of the log-spectral amplitude is subjectively a more meaningful distortion measure than the MSE of the spectral amplitude, attempts have been made to derive MMSE estimates of the log-spectral amplitude of clean speech [69, 179]. In [69], using the Gaussian statistical model as in [68], an MMSE log-spectral amplitude estimator is obtained by minimizing $\mathrm{E}\{(\log A(k) - \log \hat{A}(k))^2\}$, where $A(k)$ is the amplitude of the k^{th} spectral component of clean speech and $\hat{A}(k)$ is its estimate. The amplitude is then obtained as $\hat{A}(k) = \exp \mathrm{E}\{\log(A(k))|Y(k)\}$ and the resulting estimator is

$$H_{\mathrm{LSA}}(k) = \frac{\xi_k}{1+\xi_k} \exp\left(\frac{1}{2}\int_{v_k}^{\infty} \frac{e^{-t}}{t} dt\right). \tag{37}$$

This approach was found to result in lower residual noise than when the MSE was minimized in the spectral domain [69], which can be explained by the higher suppression provided by the LSA scheme (see Fig. 2). Suppression curves for different values of ξ_k are plotted as a function of the instantaneous SNR $\frac{|Y(k)|^2 - P_w(k)}{P_w(k)}$ (which equals $\gamma_k - 1$) in Fig. 2 for the Wiener filter, the MMSE short-time spectral amplitude estimator (STSA) in (34), the spectral power estimator (SP) in (36) and the log-spectral amplitude estimator (LSA) in (37).

It was noted in [69, 139] that incorporating the uncertainty of speech presence into the LSA estimator did not result in noticeable improvement. A multiplicatively modified LSA estimator was proposed in [139] to improve performance. Observing that a multiplicative modification of the LSA estimator to exploit the speech presence uncertainty is nonoptimal, an optimally modified estimator is derived in [38] where it is shown to provide better results. Used as a noise-reduction front-end, the LSA gain function has been shown to improve performance in the adaptive multi-rate [153] and mixed-excitation linear prediction coders [148, 150].

Bayesian estimators of the magnitude spectrum based on perceptually motivated distortion criteria are derived in [136]. Instead of the MSE in the spectral domain as in [68] or the MSE in the log-spectral domain as in [69], other measures such as the Itakura-Saito and weighted Euclidean distance are considered. Similar to the perceptual weighting used in LP based speech coders, the frequency spectrum of the error is shaped so as to place less emphasis near spectral peaks than near the valleys. This approach was found to result in better perceptual quality of the enhanced signals. Minimizing the MSE between clean and estimated spectral amplitudes each raised to the power β, is considered in [227]. For $\beta = 1$, the method reduces to the Ephraim-Malah amplitude estimator [68]. By adapting the value of β to the frame SNR, good performance is reported.

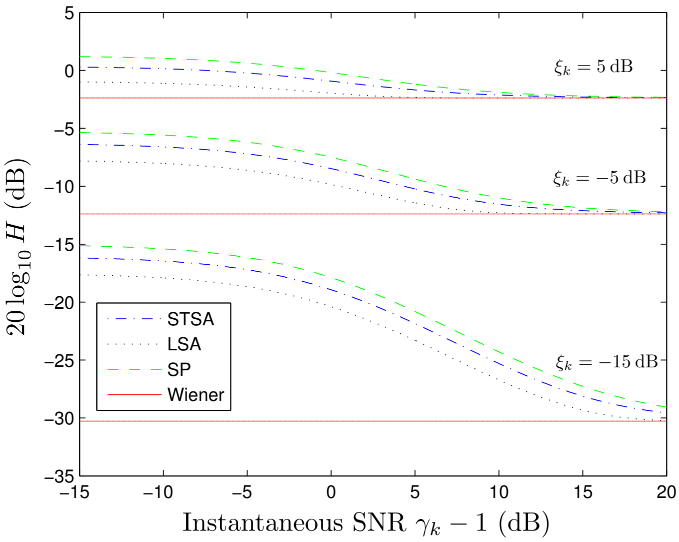

Figure 2: Suppression curves for the Wiener filter, the MMSE short-time spectral amplitude estimator (STSA) in (34), the spectral power estimator (SP) in (36) and the log-spectral amplitude estimator (LSA) in (37).

Estimating the a-priori SNR

The methods described in [68, 69, 196, 223] all need to estimate the a-priori SNR ξ_k for each frequency component k. The decision directed estimation approach discussed in [68] is one of the most commonly used techniques for this purpose. The a-priori SNR for the k^{th} frequency component and the n^{th} block is estimated as

$$\xi_k(n) = \alpha \frac{\hat{A}_k^2(n-1)}{P_{w_k}(n-1)} + (1-\alpha)Q[\gamma_k(n) - 1], \tag{38}$$

where $0 \leq \alpha < 1$ is a smoothing parameter and $Q[\cdot]$ is an operator such that $Q[x] = x$ for $x \geq 0$ and is zero otherwise.

A typical value of the smoothing parameter α is 0.98. Values of α that are close to one result in a smooth evolution of the a-priori SNR. Since the attenuation of the noisy spectral amplitude depends on the a-priori SNR, its smooth behavior eliminates large variations across successive frames, resulting in reduced musical noise [28]. This effect however comes at the cost of a slow response to an abrupt increase in the instantaneous SNR, which has an adverse effect on low energy signal components at transients and speech onsets. Taking into account the time-correlation between successive speech spectral components, causal and non-causal estimators of the a-priori SNR are derived in [41, 42]. By exploiting future data, the non-causal estimator is able to handle onsets and transients better. This approach is suitable for applications that can tolerate some amount of delay (a delay of around 100 ms is suggested in [41]). The resulting estimator is reported to preserve onsets and provide improved performance in terms of segmental SNR compared to the decision directed approach.

Super-Gaussian models

The methods discussed in the previous section assume that the speech DFT coefficients follow a Gaussian distribution. In this section, we discuss methods that assume a super-Gaussian distribution. Super-Gaussian random variables, also called leptokurtic, have a positive kurtosis. They have a more *peaky* pdf than Gaussian random variables and possess heavier tails, e.g., Laplace and Gamma distributions.

It is argued in [145] that the DFT coefficients of speech are better modelled by a Gamma distribution. Under a Gaussian assumption for speech and noise, the estimator is linear (the Wiener filter). Assuming a Gamma distribution for speech and either a Gaussian or Laplacian distribution for noise, two non-linear MMSE estimators of the complex DFT coefficients are derived. Experimental results reported in [145] show a small but consistent improvement in terms of SNR over the Wiener filter. For high a-priori SNR (e.g., 15 dB) the estimator exhibits a behavior similar to the Wiener filter. For the case when a Laplacian model is used for noise, for low a-priori SNR (e.g., -10 dB), the attenuation is constant regardless of the magnitude of the noisy DFT coefficient, resulting in reduced musical noise.

Assuming a Gamma distribution for speech and Laplace or Gaussian for noise, MMSE estimates of the squared magnitude of the speech DFT coefficients are obtained in [22]. Here too, it was observed that using a Gaussian model for the noise signal resulted in musical noise, which was avoided by the Laplace model.

Maximum a-posteriori (MAP) estimation of the spectral amplitude using super-Gaussian speech priors is presented in [137]. A parametric function is used to model the pdf of the amplitude of the k^{th} spectral component:

$$p(A_k) = \frac{\mu^{\alpha+1} A_k^{\alpha}}{\Gamma(\alpha+1)\sigma_k^{\alpha+1}} \exp\left(-\mu \frac{A_k}{\sigma_k}\right), \tag{39}$$

where $\Gamma(\cdot)$ is the Gamma function and σ_k^2 is the variance of the k^{th} component. Equation (39) approximates the Laplace distribution for $\alpha = 1$ and $\mu = 2.5$, and the Gamma distribution for $\alpha = 0.01$ and $\mu = 1.5$ [137]. The resulting estimator is computationally simpler than the super-Gaussian spectral estimator of [145].

MMSE estimation of the complex DFT coefficients under a Laplacian model for speech is discussed in [147]. The resulting estimator has a simpler analytic form compared to the case when a Gamma prior was used for speech. As in [145], using a Laplace model for noise as well results in less musical noise. MMSE and ML estimates are obtained in the DCT domain using a Laplace model for speech and a Gaussian model for noise in [82, 83], where it shown to perform better than when using a Gaussian speech prior.

A detailed theoretical and experimental analysis of MMSE estimation assuming different super-Gaussian (Laplace and Gamma) priors for speech and noise is presented in [146]. A Gaussian noise model and a super-Gaussian speech model was found to provide a higher segmental SNR than the Wiener filter, which assumes a Gaussian model for both speech and noise. Using a Laplacian noise model was found to achieve better segmental SNR only for high input-SNR conditions, but resulted in more natural residual noise. The Laplacian speech model was favored over the Gamma model as it resulted in lower musical noise. In comparison to the Ephraim-Malah amplitude estimators [68, 69], the super-Gaussian schemes achieve a higher segmental SNR but the residual noise was found to be less natural. Adaptive a-priori SNR smoothing and limiting [148] are suggested for improving the quality [146].

3.4 Trained statistical model based systems

The methods discussed in the previous section are optimal only within the framework of the statistical models they assume. Rather than describing complex signals such as speech with models with few parameters, a more accurate method is to use more sophisticated statistical models such as HMMs, GMMs and codebooks that have been trained using a representative database. The improved accuracy is at the expense of a higher computational complexity compared to methods such as [68, 146].

In trained model based speech enhancement, the pdfs of the speech and noise processes are estimated from corresponding training sequences. To simplify the estimation, the processes are described by parametric models (e.g., the AR model), whose parameters are then estimated from the data. The theoretical analysis in the training and use of models such as GMMs or codebooks requires that the signals are stationary. In practice, to deal with the nonstationarity of the speech signal, processing is performed in blocks of 20 - 30 ms within which the signal can be assumed to be stationary, and by allowing the pdf to have multiple modes (e.g., using GMMs or codebooks). These models are then used to obtain either MMSE or MAP estimates of the speech signal.

First we discuss existing HMM based enhancement schemes, e.g., [65, 66, 188]. Next we consider the codebook based approaches described in papers A and B, and describe how they address the shortcomings of the above mentioned HMM schemes in nonstationary noise. Finally, differences between the two approaches are summarized in section 3.4.

HMM based methods

HMMs have been used extensively in speech recognition [181, 182]. In [65–67], HMMs trained on clean speech and noise were used for speech enhancement, and Bayesian MMSE and MAP estimates of the speech signal were obtained. The HMMs consist of several states with a mixture of Gaussian pdfs at each state. A state transition matrix governs the transition from one state to another[6]. The covariance matrix of each Gaussian pdf is parameterized by the AR parameters of the signal. The AR parameters are the linear predictive coefficients and the variance of the excitation signal.

In the MAP approach, an estimate of the speech signal is obtained by maximizing the posterior pdf of the speech signal given the noisy observations. Since the corresponding gradient equations are nonlinear, a local maximization is performed using the EM algorithm. In the MMSE approach, a weight is associated with the Wiener filter corresponding to each combination of speech and noise components at each state. The MMSE estimate of the clean speech signal is obtained by filtering the noisy signal with the weighted sum of these Wiener filters over all combinations of states and mixtures.

As mentioned earlier, the HMM models both the LP coefficients and the excitation variance (gain) [66]. This generally leads to a mismatch in the gain term between training and testing. Thus some form of gain adaptation is essential. For the MAP estimation described in [66], gain-normalized HMMs are trained for the clean speech signal. Let $\lambda = (\lambda_x, \lambda_w)$, where λ_x denotes the parameter set for the gain-normalized HMM for the clean signal and λ_w denotes the parameter set for the noise HMM. First, the gain of the noise model is adjusted based on an estimate of the noise statistics made from the noisy observation. At time instant t, gain-adapted MAP signal estimation is then performed according to

$$\hat{\mathbf{x}}_t = \max_{\mathbf{x}_t} \max_{g_t > 0} p_\lambda(\mathbf{x}_t, \mathbf{y}_1^t | g_1^t), \tag{40}$$

where g_t is the gain corresponding to the current frame at time t, $\mathbf{x}_t$ is the vector of clean speech samples at time t (corresponding to a single frame), $\mathbf{y}_1^t$ is the sequence of vectors of noisy samples up to time t, g_1^t is the gain contour of the speech model and $p_\lambda(\mathbf{x}_t, \mathbf{y}_1^t | g_1^t)$ is the joint pdf of $\mathbf{x}_t$ and $\mathbf{y}_1^t$, given the gain contour g_1^t and the complete parameter set λ.

It is important to note that g_t is optimized based on the noisy observation and the parameter set of the noise model. In [66, 188], the noise model

[6] Alternatively, a GMM, which can be seen as a single-state HMM, can also be used. GMM based speech enhancement methods are presented in [7, 8, 27, 168, 232, 233].

is obtained during speech pauses. In [135], the speech HMM (which includes a model for silence) is used to detect pauses. For stationary noise, using the entire speech utterance, ML estimation of the noise model parameters through the EM approach is also proposed [135, 187]. One straightforward modification is to use the more accurate noise estimates provided by the minimum statistics approach [144], which we found resulted in better performance than estimation based on speech pauses. However, since the gain adaptation adapts to changing levels of background noise only during the next speech pause or only as quickly as the buffer length in the long-term noise estimation [144] allows, it still suffers from poor performance in highly nonstationary noise.

Codebook based approach

The codebook based approaches [125,206, paperA, paperB] attempt to overcome the disadvantage of the HMM methods discussed above in nonstationary noise. An instantaneous frame-by-frame gain computation was introduced in [125] and extended in [206, paper A]. Such an approach was also considered in a speech decomposition context in [211]. In [125, paper A, paper B], using trained codebooks of only the LP coefficients of speech and noise, the gain terms are computed for each short-time frame based on the LP coefficients and the noisy observation. The codebooks are trained using representative databases of speech and noise.

In paper A, which describes a maximum likelihood approach, the speech and noise codebook indices and the excitation variances corresponding to the vectors that the indices represent are obtained according to:

$$\{i^*, j^*, {\sigma_x^2}^*, {\sigma_w^2}^*\} = \arg\max_{i,j,\sigma_x^2,\sigma_w^2} p(\mathbf{y}|\theta_x^i, \theta_w^j, \sigma_x^2, \sigma_w^2), \tag{41}$$

where σ_x^2 and σ_w^2 are the excitation variances of clean speech and noise respectively, and $\theta_x^i = (a_{x_0}^i, \ldots, a_{x_p}^i)$ and $\theta_w^j = (a_{w_0}^j, \ldots, a_{w_q}^j)$ are the LP coefficients of clean speech and noise with p and q being the respective LP-model orders. A schematic diagram of this method is shown in Fig. 3. Using the equivalence between the log-likelihood and the Itakura-Saito distortion [111], the estimation can be performed in the frequency domain according to

$$\{i^*, j^*\} = \arg\min_{i,j} \left\{ \min_{\sigma_x^2,\sigma_w^2} d_{\mathrm{IS}} \left(P_y(\omega), \frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2} \right) \right\}, \tag{42}$$

where

$$d_{\mathrm{IS}}(P_y, \hat{P}_y) = \frac{1}{2\pi} \int_0^{2\pi} \left(\frac{P_y(\omega)}{\hat{P}_y(\omega)} - \ln(\frac{P_y(\omega)}{\hat{P}_y(\omega)}) - 1 \right) d\omega \tag{43}$$

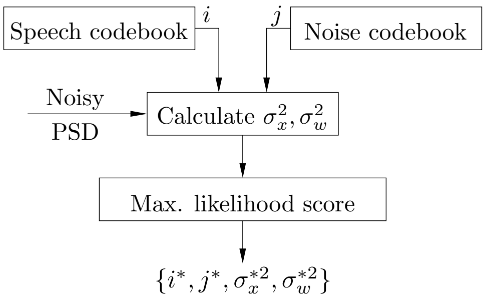
Speech codebook
i
j
Noise codebook
Noisy
PSD
Calculate σ_x^2, σ_w^2
Max. likelihood score
$\{i^*, j^*, \sigma_x^{*2}, \sigma_w^{*2}\}$

Figure 3: Estimation of excitation variances and spectral shapes: i^*, j^* are the indices of the selected entries from the speech and noise codebooks and $\sigma_x^{*2}, \sigma_w^{*2}$ are the corresponding excitation variances.

and

$$A_x^i(\omega) = \sum_{k=0}^{p} a_{x_k}^i e^{-j\omega k}, \quad A_w^j(\omega) = \sum_{k=0}^{q} a_{w_k}^j e^{-j\omega k}. \tag{44}$$

For given $A_x(\omega)$ and $A_w(\omega)$, the excitation variances that minimize the Itakura-Saito distortion in (42) can be obtained under the assumption of small modeling errors by using a series expansion for $\ln(x)$ up to second order terms [206, eqn. 11].

The ML estimate of the codebook vectors and the variances can be used in applications that require estimates of the clean speech AR parameters. For example, a Wiener filter can be constructed according to

$$H(\omega) = \frac{\dfrac{\sigma_x^{2^*}}{|A_x^{i^*}(\omega)|^2}}{\dfrac{\sigma_x^{2^*}}{|A_x^{i^*}(\omega)|^2} + \dfrac{\sigma_w^{2\,*}}{|A_w^{j^*}(\omega)|^2}}, \tag{45}$$

which can then be used to estimate the clean signal from the noisy observation.

As another example, the ML estimation approach can be viewed as a modified distortion measure used to select an entry from the speech LP codebook under noisy conditions, a feature that can be easily integrated into parametric coders that require accurate estimates of the spectrum [207]. A sufficiently large speech codebook is necessary to provide an acceptable accuracy in the parameter description. Multi-stage speech codebooks can be used for this purpose and the indices resulting from the ML search can be transmitted to the decoder. At each stage, we choose the codebook entry that results in the highest likelihood. A configuration with a two stage speech codebook is shown in Fig. 4. The ML search using the first stage results in the selection of a single speech codebook entry as the ML estimate. The second stage speech codebook forms an additive refinement to this codebook entry, producing a refined codebook. The ML search is repeated with the refined codebook. The two resulting indices (one for each stage) can be transmitted to the decoder. The search can be generalized in a straightforward manner to more than two stages.

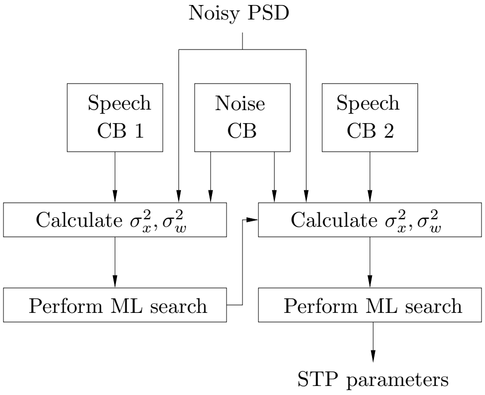

Figure 4: VQ search under noisy conditions using a two stage speech codebook. The second codebook forms an additive refinement to the speech LP vector resulting from the ML search of the first stage.

In a practical implementation, multiple noise codebooks are used, each trained on a different noise type. For each segment of noisy speech, a classification is made using this long-term estimate and a particular noise codebook is selected. The selected noise codebook is then used in the subsequent ML search. One way to perform the classification is using the spectral shape obtained from the long-term noise estimates provided by [144]. We note that though the classification is performed using long-term noise estimates, the different entries in the codebook permit variations in the spectral shape. Moreover, the optimal gain computation is still performed instantaneously. What the classified scheme cannot handle instantaneously is when the pdf (considering only the spectral shape) of the noise varies rapidly since the long-term estimate cannot adapt immediately. However, in most practical situations, the noise distribution does not change rapidly.

In paper B, a codebook based MMSE estimation of the speech and noise AR parameters with frame-by-frame gain computation is proposed. While in the ML approach of paper A, one pair of speech and noise LP vectors was selected as the ML estimate, the MMSE estimate of the speech (noise) vector is a weighted sum[7] of the speech (noise) codebook vectors. Similarly, the MMSE estimate of the speech and noise excitation variances is the weighted sum of the excitation variances corresponding to each pair of speech and noise codebook vectors, and the current frame of the noisy observation. Thus, the MMSE estimation can be seen as a soft-decision procedure that allows for a proportionate contribution from vectors according to their probability given the observation. Both memoryless (using information from the current frame alone) and memory-based (using information from the current and previous frames) estimators are derived. Estimation of functions of the speech and noise AR parameters is also addressed, in particular one that leads to the MMSE estimate of the clean speech signal. The codebook based MMSE estimator takes into account the a-priori probabilities of each of the speech and noise codebook vectors.

The speech pdf

Let θ_x and θ_w denote the random variables corresponding to the speech and noise LP coefficients respectively. Let σ_x^2 and σ_w^2 denote the random variables corresponding to the speech and noise excitation variances respectively. In the codebook-based method[8], the speech pdf can be written as

$$p(\mathbf{y}) = \int_{\Theta} p(\mathbf{y}|\theta)p(\theta)d\theta, \tag{46}$$

where $\theta = [\theta_x,\, \theta_w,\, \sigma_x^2,\, \sigma_w^2]$. The integral is over $\Theta = \Theta_x \times \Theta_w \times \Sigma_x \times \Sigma_w$, where Θ_x and Θ_w represent the support-space of the vectors of speech and noise LP coefficients and Σ_x and Σ_w represent the support-space for the speech and noise excitation variances. Note that θ specifies the covariance matrix of $\mathbf{y}$. We assume that the conditional pdf $p(\mathbf{y}|\theta)$ is Gaussian. However, the marginal pdf $p(\mathbf{y})$ can be seen from (46) to be a mixture of Gaussians. This, and the fact that a data-driven codebook is used, make the resulting model more flexible than assuming a Laplacian or Gaussian marginal.

Differences between the HMM and codebook approaches

The main difference between the HMM methods described in [66, 67, 188] and the codebook approaches of papers A and B lies in the manner in which they handle the nonstationarity of the noise signal, which in turn is related to the modelling and computation of the excitation variances. Since the HMM method models both the LP coefficients and the excitation variance as prior information [67], a gain adaptation is required for the

[7] The weighted addition is performed in the LSF domain.

[8] A similar expression can be obtained for the HMM based methods as well. For simplicity, we consider only the codebook method in this discussion.

speech and noise models to compensate for differences in the level of the excitation variance between training and operation. The gain adaptation factor is computed using the observed noisy gain and an estimate of the noise statistics obtained using, e.g., the minimum statistics approach [144]. Conventional noise estimation techniques are buffer-based techniques, where an estimate is obtained based on a buffer of several past frames. Thus, such a scheme cannot react quickly to nonstationary noise. In the codebook based approach, the codebook models only the LP coefficients, and the speech and noise excitation variances are optimally computed on a frame-by-frame basis, using the noisy observation. This enables the method to react quickly to nonstationary noise. We note that recently, motivated by the frame-by-frame gain computation of the codebook based methods [125, paper A, paper B], an HMM based enhancement scheme with explicit noise gain modelling and on-line estimation has been proposed [230].

Another difference is that the HMM based methods of [66,67,188] obtain MMSE estimates of the clean speech signal whereas the codebook approach obtains MMSE estimates of the speech and noise STP parameters. Let the vector $\mathbf{X}$ denote the random variable corresponding to a frame of the clean speech signal. Given the noisy observations, the HMM method obtains the expected value of $\mathbf{X}$ and its functions such as the spectral magnitude and the log-spectral magnitude. The codebook method obtains the expected value of θ given the noisy observations for the current and previous frames, which is useful in applications that require optimal estimates of the speech and noise AR parameters. The framework developed in the codebook approach also allows the MMSE estimation of functions of the speech and noise AR parameters, where the MMSE estimate of one such function can be shown to result in the expected value of $\mathbf{X}$ given the noisy observations [209], which is useful in applications where an optimal estimate of the time domain speech waveform is desired.

Computational complexity

As discussed earlier, the HMM based methods and the codebook based approaches employ a more accurate model for the speech pdf compared to the methods of sections 3.2 and 3.3. The price to be paid for the improved accuracy is an increase in computational complexity. The complexity is directly related to the model size, e.g., the number of codebook vectors, or the number of states and mixture components in the HMM. In paper A, an iterative scheme to reduce computational complexity is proposed. It is also relevant to mention that the HMM and codebook approaches lend themselves in a straightforward fashion to parallel processing, which can result in a significant speedup. For example, in the ML approach of paper A, in principle, one processor can be assigned to compute the likelihood $p(\mathbf{y}|\theta_x^i, \theta_w^j, \sigma_x^2, \sigma_w^2)$ corresponding to each combination of speech and noise codebook vectors.

The amount of time required for the resulting computations is independent of the model size[9]. An additional step of weighted summation is required for the MMSE approaches, though the computation of the likelihood can still be performed in parallel.

3.5 Complements

Given the vast and diverse literature on speech enhancement, a categorization of methods as above cannot be comprehensive. In this section, for completeness, we briefly mention and provide references to a subset of other enhancement schemes.

In [132], the speech signal is modelled as the response of an all-pole system using a Gaussian AR model, and the MAP estimate of the speech signal and its AR parameters given the noisy speech is derived. The resulting equations for the joint MAP estimation are non-linear and a sub-optimal iterative solution is proposed. The AR parameters at a particular iteration are estimated from the estimate of the clean signal at that iteration. The clean signal is then re-estimated using the new AR parameters and the iteration continues. This method was investigated in [98] and was found to suffer from some drawbacks, e.g., no proper convergence criterion was defined, and the formant bandwidths decreased with increasing number of iterations. Inter-frame and intra-frame constraints were introduced to ensure the stability of the all-pole model, to ensure that the AR parameters were speech-like and not to allow a high variation of the parameters in successive frames [98, 176]. The optimum number of iterations from a perceptual sense were determined empirically for different classes of speech sounds. The constrained iterative procedure was further improved in [201, 202] by constraining the estimated AR parameters to belong to a trained codebook of speech AR coefficients.

The sinusoidal model representation of speech signals [157, 158] has also been employed in enhancement. In [4], a Wiener filter is applied to the multi-resolution sinusoidal transform parameters, which are reportedly well matched to the human auditory system [3]. A constrained iterative sinusoidal model is employed in [113]. The sinusoidal amplitudes are estimated in an iterative fashion, as a weighted average of the estimate from the previous iteration and its Wiener-filtered counterpart. At a given iteration, the Wiener filter is constructed using the current estimate of the amplitude and an estimated noise amplitude. Further, it is ensured that the amplitudes evolve smoothly over time. The number of iterations was empirically determined to be seven. In voiced regions, a smoothing procedure is applied

[9] This is an extreme case. In general, a speedup is guaranteed with the use of more than one processor, and the resulting computational complexity is determined by the model size and the number of processors.

to the sinusoidal frequencies. The phase is not modified. The method was observed to perform well in voiced regions.

Most speech signal processing algorithms require wide-sense stationarity assumptions on the signal. However, it is well known that speech can at best be described as a quasi-stationary signal. As a result, most speech processing algorithms divide the speech signal into small, fixed-size segments within which stationarity can be assumed to be preserved. The trade-off between the resolution in time and in frequency is a natural by-product of such a scheme. In [203], an adaptive segmentation scheme that divides the signal into the longest segments within which stationarity is preserved [140] is shown to result in reduced musical noise. It is also possible to obtain improved estimates of the noisy power spectra (or the noisy covariance matrix) for use in enhancement by employing an adaptive segmentation [100, 101].

Following the successful exploitation of auditory masking in audio coding [116], attempts have been made to apply psychoacoustic principles to speech enhancement as well. Some of these methods have already been mentioned in connection with spectral subtraction and the subspace methods in section 3.2. One of the early enhancement approaches to exploit auditory properties used the concept of lateral inhibition [33, 34]. Noise components that lie below the masking threshold can be left unchanged, resulting in lower speech distortion. This is achieved in [212, 213] by defining and estimating an audible noise component, which is then suppressed. In [95], instead of a complete removal of the (audible) noise, a noise-floor is defined and the masking threshold is used to design a weighting rule to ensure that the perceived noise suppression equals a pre-defined level. This results in an enhanced signal with a residual noise that sounds natural. This approach was extended to joint acoustic echo cancellation and noise reduction for hands-free systems in [96].

When using a state-space representation to formulate the speech enhancement problem in a sequential estimation framework, closed form analytical solutions are available only in certain special cases, e.g., using a linear Gaussian state-space representation results in the Kalman filter [85]. For more general (e.g., non-linear) state-space models, approximate methods are used, e.g., the extended Kalman filter [2]. An alternate strategy, reported to result in better performance, is to use Monte Carlo integration (also known as particle methods) [5, 62]. Particle filtering and smoothing have been applied to speech enhancement, for the white noise case [73,88,215]. The noise variance is assumed known or estimated from speech pauses. These methods consider a time-varying AR model for the speech signal. An advantage of the Monte Carlo approach is that the estimation accuracy is independent of the dimension of the state-space and only depends on the number of particles.

4 Multi-channel speech enhancement

Multi-channel enhancement algorithms exploit the spatial diversity resulting from the fact that the desired and interfering signal sources are in practice located at different points in space. This diversity can be taken advantage of, e.g., by steering a null towards the noise source and a beam towards the signal source.

Microphone array based noise reduction is useful in a variety of applications such as hands-free communication inside a car [64, 72, 93, 143, 152, 171], in tele-conferencing, and as a front-end for speech recognition [15, 87, 160, 161, 167, 169, 170, 189]. Another application is in hearing aids [120, 133, 177, 198–200, 222], where some of the latest models feature up to three microphones.

In this section, we provide a brief overview of some common multi-channel noise reduction techniques. We begin with introducing the signal model, followed by a description of beamforming techniques, multi-channel Wiener filtering and a discussion on blind source separation.

4.1 Signal model

We assume a far-field model so that wave propagation can be assumed to be planar. The signals arriving at the different sensors are attenuated equally and differ only in their phase (they are delayed versions of one another). The different sensor signals can be assumed to have identical power spectra, since, in practice, for closely spaced sensors, the delay between the sensors is very small compared to the short-time stationarity of the speech signal.

Consider a speech source located at an angle θ from the array. Let d_i denote the distance of the i^{th} sensor from the center of the array, x_0. We assume a fixed inter-element spacing, i.e., $d_i - d_{i-1} = d$. The additive noise model can then be written in the frequency domain as

$$\tilde{\mathbf{Y}}(k) = X(k)\mathbf{d} + \tilde{\mathbf{W}}(k), \tag{47}$$

where $\tilde{\mathbf{Y}}(k) = [\tilde{Y}_1(k) \dots \tilde{Y}_M(k)]^T$, $\tilde{\mathbf{W}}(k) = [\tilde{W}_1(k) \dots \tilde{W}_M(k)]^T$, $\tilde{Y}_i(k)$ is the noisy signal observed at the i^{th} sensor, $X(k)$ corresponds to the clean speech component at the center of the array, $\mathbf{d} = [e^{-j\omega\tau_1} \dots e^{-j\omega\tau_M}]^T$ is referred to as the steering vector, $\tilde{W}_i(k)$ corresponds to the background noise at the i^{th} sensor, and k is the discrete frequency index. The delay τ_i at the i^{th} sensor, relative to the center of the array, is given in samples according to

$$\tau_i = \frac{d_i \cos\,\theta}{c} f_s, \tag{48}$$

where f_s is the sampling frequency and $c = 340$ m/s is the speed of sound in air. We assume that the array has been steered towards the speech source, which can be achieved by compensating for the relative delays [57, 124] as shown in Fig. 5. This results in

$$\mathbf{Y}(k) = X(k)\mathbf{1} + \mathbf{W}(k), \tag{49}$$

where $\mathbf{Y}(k) = [Y_1(k) \ldots Y_M(k)]^T$, $\mathbf{W}(k) = [W_1(k) \ldots W_M(k)]^T$, $Y_i(k)$ and $W_i(k)$ are the noisy and noise signal components corresponding to the i^{th} sensor after steering, and $\mathbf{1}$ is a $M \times 1$ vector of ones.

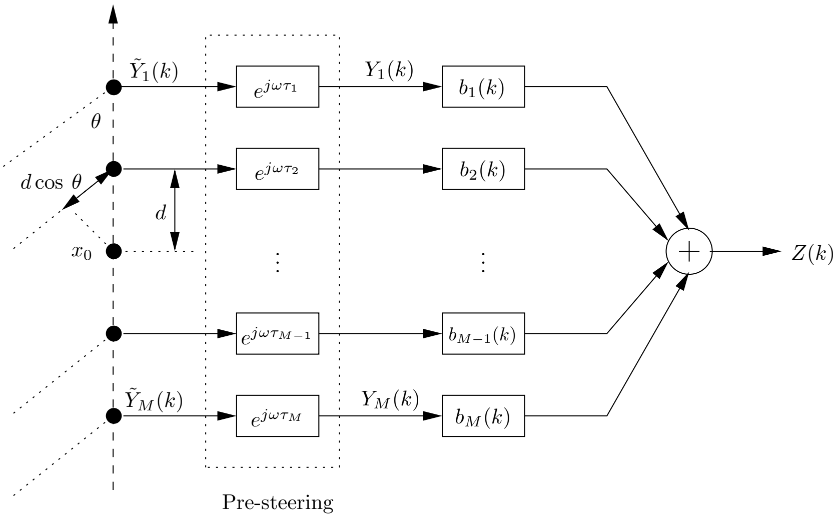

Figure 5: Microphone array steering and beamforming assuming a far-field model. The signal is incident on the array at an angle θ. The inter-element spacing is d. $\tilde{Y}_i(k)$ and $Y_i(k)$ are the noisy signal components corresponding to the i^{th} sensor before and after steering. $b_1(k), \ldots, b_m(k)$ are the beamformer weights and $Z(k)$ is the output of the beamformer. In this figure, $\omega = \frac{2\pi k}{K}$. The figure is adapted from [56].

4.2 Beamforming

Beamforming is a means of performing spatial filtering [214]. In the frequency domain, beamforming can be viewed as a linear combination of the sensor outputs:

$$Z(k) = \sum_{i=1}^{M} b_i(k) Y_i(k), \tag{50}$$

$b_i(k)$ is the beamformer weight corresponding to the i^{th} sensor, and M is the total number of sensors. In vector notation, we have

$$Z(k) = \mathbf{b}^T(k)\mathbf{Y}(k), \tag{51}$$

where $\mathbf{b}(k) = [b_1(k) \dots b_M(k)]^T$. Beamforming can be classified into two categories - fixed, where the weights are fixed across time, and adaptive, where the weights vary in response to changes in the acoustic environment.

Fixed beamforming

In fixed beamforming, the weights $b_i(k)$ are fixed over time, and are determined by minimizing the power of the signal at the output of the beamformer subject to a constraint that ensures that the desired signal is undistorted [12], i.e., the optimal weights are the solution to

$$\min_{\mathbf{b}(k)} \mathbf{b}^*(k)\Phi_{yy}(k)\mathbf{b}(k) \qquad \text{subject to} \qquad \mathbf{b}^*(k)\mathbf{1} = 1, \tag{52}$$

where $*$ refers to complex conjugate transpose and $\Phi_{yy}(k)$ is the $M \times M$ PSD matrix of the noisy input signals whose $(i,j)^{th}$ entry is $\mathrm{E}[Y_i(k)Y_j^*(k)]$. Note that the constraint of zero distortion in the look direction is written using a vector of ones since we assume that the array has been pre-steered towards the desired signal direction. The solution to the constrained optimization problem (52) is the well-known minimum variance distortionless response (MVDR) beamformer [49]:

$$\mathbf{b}(k) = \frac{\Phi_{ww}^{-1}(k)\mathbf{1}}{\mathbf{1}^T\Phi_{ww}(k)\mathbf{1}}, \tag{53}$$

where $\Phi_{ww}(k)$ is the $M \times M$ noise PSD matrix whose $(i,j)^{th}$ entry is $\mathrm{E}[W_i(k)W_j^*(k)]$. Assuming a homogeneous noise field, the solution can be written in terms of the coherence matrix

$$\mathbf{b}(k) = \frac{\Gamma_{ww}^{-1}(k)\mathbf{1}}{\mathbf{1}^T\Gamma_{ww}(k)\mathbf{1}}, \tag{54}$$

where the $(i,j)^{th}$ entry of the $M \times M$ coherence matrix is given by

$$\begin{aligned}\Gamma_{ij}(k) &= \frac{\phi_{w_iw_j}(k)}{\sqrt{\phi_{w_iw_i}(k)\phi_{w_jw_j}(k)}} &(55)\\ &= \frac{\phi_{w_iw_j}(k)}{\phi_{ww}(k)}, &(56)\end{aligned}$$

where $\phi_{w_iw_j}(k)$ is the cross spectral density between the noise signals at the i^{th} and j^{th} sensors, and from the assumption of a homogeneous noise field,

$\phi_{w_iw_i}(k) = \phi_{ww}(k)$ for all i. A schematic diagram of the fixed beamformer is shown in Fig. 5.

For incoherent (or spatially white) noise fields, $\Gamma_{ww} = I$, $\mathbf{b} = \frac{1}{M}\mathbf{1}$ and the MVDR beamformer reduces to a delay-and-sum beamformer (DSB), where the sensor signals are delayed and then averaged. The pre-steering corresponds to the delay and is such that the signal components at the different sensors sum up constructively while the noise components cancel each other. Incoherent noise fields are not common. An example of incoherent noise is electrical noise at the sensors, which is uncorrelated at the different sensors.

In a DSB, the amplitude weights are fixed across frequency (often equal) and the phase weights introduce the delay. A more general form is a filter-and-sum beamformer (FSB), where both the amplitude and phase weights vary across frequency. FSBs are useful in designing beamformers with a specified directivity pattern for arbitrary microphone array configurations [58, 119].

Many of the noise fields encountered in practice fall into the category of diffuse noise fields, whose coherence function has the form [12]:

$$\Gamma_{ij}(k) = \operatorname{sinc}\left(\frac{2\pi k}{K}\frac{d_{ij}}{c}\right), \tag{57}$$

where $\operatorname{sinc}(x) = \sin(x)/x$, d_{ij} is the distance (in meters) between the i^{th} and j^{th} sensors, $c = 340$ m/s is the speed of sound in air and K is the frame length. If we use the corresponding expression for the coherence matrix in (54), the resulting beamformer is called a superdirective beamformer (SDB) [61, 151]. While the SDB is useful in diffuse noise fields, its main disadvantage is an amplification of uncorrelated noise (e.g., sensor noise) at low frequencies. This problem is handled by incorporating a white noise gain constraint in the design [11, 49, 86].

Adaptive beamforming

In adaptive beamforming, the beamformer weights adapt to changes in the acoustic environment over time. The optimal weights are obtained by minimizing the variance of the output signal. To ensure that the speech signal is not cancelled out or distorted, a distortionless constraint is imposed on the desired signal. This results in the linearly constrained minimum variance (LCMV) beamformer [75], where the adaptive beamformer weights are obtained through a constrained minimization procedure.

The generalized sidelobe canceller (GSC) [14, 21, 25, 26, 81, 94, 172, 199], is an efficient alternative implementation of Frost's LCMV approach, that converts the constrained optimization problem into an unconstrained one. This leads to an efficient implementation for the update of the beamformer weights.

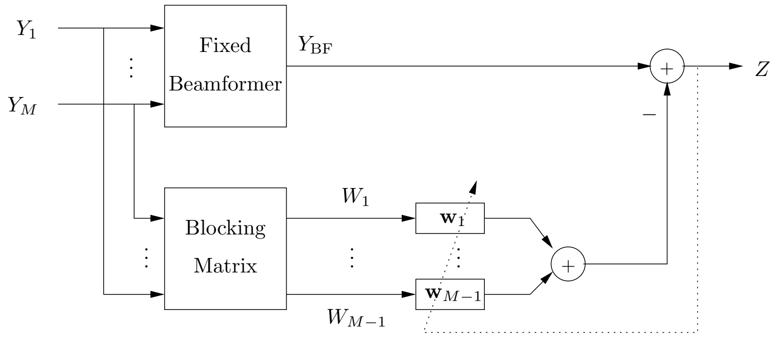

Figure 6: Frequency domain implementation of the Generalized Sidelobe Canceller. The ANC is implemented by the adaptive filters $\mathbf{w}_1, \ldots, \mathbf{w}_{M-1}$.

The GSC consists of three parts - a fixed beamformer (FBF), a blocking matrix (BM) and an adaptive noise canceller (ANC) as shown in Fig. 6. The FBF includes a pre-steering module and its weights are designed to produce a speech reference Y_{BF} with a specified gain and phase response. The FBF could either be a simple delay-and-sum beamformer, or a more advanced filter-and-sum or superdirective beamformer. The BM is generally orthogonal to the FBF and produces $M-1$ outputs, called the noise references, by steering zeros towards the desired signal direction. One way to create the noise references is to take the difference between adjacent sensor signals [94]. The ANC (implemented by the adaptive filters $\mathbf{w}_1, \ldots, \mathbf{w}_{M-1}$ in Fig. 6) removes any remaining correlation between the speech reference Y_{BF} and the noise references. Thus, any residual noise in the speech reference that is correlated to the noise references is removed.

In practice, the noise references are not completely free of speech. As a consequence, the ANC results in some of the speech signal being cancelled. To minimize the effect of the speech leakage on the ANC, the noise-cancelling filters are adapted only during periods of speech absence. To reduce the amount of speech leakage, some variants of the GSC employ an adaptive blocking matrix [102,104,105]. Variations of the GSC designed to improve performance in reverberant environments are presented in [80,81].

4.3 Multi-channel Wiener filtering

It can be shown that the MVDR beamformer is the optimal solution in an ML sense (assuming the noise to be Gaussian) and also the SNR-optimal solution for narrowband signals [166,192]. In this case, since the MVDR beamformer is data independent, it has an advantage over the multi-channel

Wiener filter (MWF), which is data dependent. However, for broadband signals such as speech, the MWF is the optimal solution in the MSE sense. The MWF can be factored into an MVDR beamformer followed by a single-channel Wiener post-filter [192]:

$$\mathbf{H}(k) = \underbrace{\frac{\phi_{xx}(k)}{\phi_{xx}(k) + (\mathbf{1}^T \Phi_{ww}^{-1} \mathbf{1})^{-1}}}_{\text{Post-filter}} \underbrace{\frac{\Phi_{ww}^{-1}\mathbf{1}}{\mathbf{1}^T \Phi_{ww} \mathbf{1}}}_{\text{MVDR}}, \tag{58}$$

where $\phi_{xx}(k)$ is the PSD of the clean speech signal. As before, we assume that the array has been pre-steered towards the speech source. The PSD of the noise after the MVDR beamforming can be shown to be equal to $(\mathbf{1}^T \Phi_{ww}^{-1} \mathbf{1})^{-1}$ so that the post-filter in (58) is in fact the Wiener filter. The post-filter is particularly advantageous in diffuse noise environments where the MVDR beamformer is not very effective.

To perform the post-filtering, estimation of the speech PSD $\phi_{xx}(k)$ is crucial and several approaches exist [13,18,40,71,141,159,162,193,228,229]. The denominator of the post-filter expression in (58) is simply the PSD of the MVDR beamformer output. The Zelinski post-filter [228,229] assumes that the background noise is uncorrelated at the different sensors. Under the assumption that the signal and noise are uncorrelated, the cross-spectral density of the microphone signals then provides an estimate of $\phi_{xx}(k)$:

$$\begin{aligned} \mathrm{E}[Y_i Y_j^*] &= \mathrm{E}[XX^*] + \mathrm{E}[W_i W_j^*] + \mathrm{E}[X W_j^*] + \mathrm{E}[W_i X^*] \\ &= \phi_{xx}. \end{aligned} \tag{59}$$

By averaging over all possible combinations of the sensors, the estimator can be made robust. A block diagram of the MWF is shown in Fig. 7.

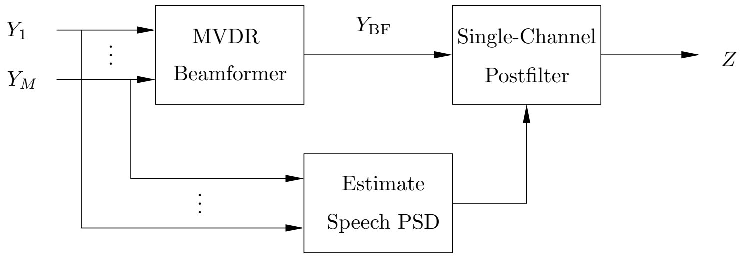

Figure 7: The multi-channel Wiener filter.

For many of the practical noise environments encountered in noise reduction applications such as inside a car or an office, the noise field is diffuse. A post-filter that accounts for diffuse noise is described in [159]. Under the assumption of a homogeneous diffuse noise field, we have

$$\begin{aligned}\phi_{y_iy_i} &= \phi_{xx} + \phi_{ww} \quad \text{for all } i \\ \phi_{y_iy_j} &= \phi_{xx} + \Gamma_{ij}\phi_{ww},\end{aligned} \tag{60}$$

where the coherence function Γ_{ij} is defined by (55) and is given by (57) for diffuse noise. An estimate of the clean speech PSD can then be obtained from the signals at sensors i and j according to

$$\hat{\phi}_{xx}^{ij} = \frac{\phi_{y_iy_j} - \frac{1}{2}(\phi_{y_iy_i} + \phi_{y_jy_j})}{1 - \Gamma_{ij}}, \tag{61}$$

where the average over $\phi_{y_iy_i}$ and $\phi_{y_jy_j}$ is taken for robustness. As before, by averaging $\hat{\phi}_{xx}^{ij}$ over all combinations of sensors, robustness can be improved.

In paper C, a parametric model-based approach for multi-channel Wiener filtering is presented. By employing an AR model for the speech signal, and using a trained codebook of speech LP coefficients, an MMSE estimate of the clean speech signal is obtained. By explicitly accounting for steering errors in the signal model, robust estimates are obtained.

4.4 Blind source separation

The goal of blind source separation (BSS) is to recover a set of independent sources given only a set of sensor observations that are generated from the individual source signals through an unknown linear mixing process. The task is blind since there is no knowledge available about either the sources or the mixing process, except that the sources are independent and that the mixing is linear. In the following, we restrict our discussion to the case when the number of sensors equals the number of sources and consider the 2×2 case for simplicity.

Let $\mathbf{x}(t) = [s_1(t) \quad s_2(t)]^T$ denote the vector of source signals and $\mathbf{y}(t) = [y_1(t) \quad y_2(t)]^T$ denote the observed sensor signals sampled at time instant t. In instantaneous BSS, the sensor signals are related to the sources according to

$$\mathbf{y}(t) = A\mathbf{x}(t), \tag{62}$$

where A is the 2×2 mixing matrix. To avoid problems associated with the inversion of an estimate of A to achieve separation, a common approach is to estimate a backward model in terms of the separating matrix W such that the output

$$\mathbf{z}(t) = W\mathbf{y}(t), \tag{63}$$

is an estimate of the source vector $\mathbf{x}(t)$ up to an arbitrary permutation and scaling. An estimate of W is generally obtained by optimizing a so-called contrast function, which is a function of the distribution of $\mathbf{z}(t)$. For example, W can be estimated by minimizing the Kullback-Leibler divergence between the pdf of $\mathbf{z}(t)$ (which is the distribution at the output of the separation) and the product of the pdfs of $s_i(t)$ expressed using W and $\mathbf{y}(t)$ (which is the source distribution according to the independence assumption) [9, 30, 31, 63]. Other examples include contrast functions based on higher-order (larger than two) cumulants [29, 31, 46], which require that at most one of the sources is Gaussian. For nonstationary source signals, it is possible to achieve separation by exploiting only second-order statistics. For example, the methods described in [48, 84, 175, 219] achieve separation through decorrelation of the signals at the output of the demixing system at multiple time instants.

In practice, for acoustic signals, the mixing is better described by a convolutive model as given by the following $Q+1$ tap mixing system:

$$\mathbf{y}(t) = \sum_{q=0}^{Q} A_q \mathbf{x}(t-q), \tag{64}$$

where A_q is a 2×2 matrix for each q. Applying the DFT to a time domain segment of length T, we have

$$\mathbf{Y}(\omega) = A(\omega)\mathbf{X}(\omega), \tag{65}$$

where $\mathbf{Y}(\omega) = [Y_1(\omega)\ Y_2(\omega)]^T$, $\mathbf{X}(\omega) = [S_1(\omega)\ S_2(\omega)]^T$ and $Y_i(\omega) = \sum_{t=0}^{T-1} y_i(t)e^{-j\omega t}, i = 1, 2$. $S_1(\omega)$ and $S_2(\omega)$ are obtained similarly from $s_1(t)$ and $s_2(t)$ respectively. $A(\omega)$ corresponds to the frequency response of the mixing filters A_q, obtained through a component-wise DFT. The BSS task reduces to the estimation of an unmixing matrix $W(\omega)$ for each ω. This corresponds to the so-called narrowband approach of BSS, since a separate BSS problem is solved for each frequency bin, implying a narrowband signal model. This approach is computationally simple but suffers from a permutation and scaling problem in each frequency bin that need to be resolved in a consistent manner. To avoid this problem, broadband approaches to convolutive BSS have been proposed, where the frequency bins are no longer treated independently [23, 24].

For point noise sources, as shown in Fig. 8, the acoustic background noise reduction problem can be cast in a BSS framework where $s_1(t) = s(t)$ is the speech signal and $s_2(t) = n(t)$ is the noise signal. The decorrelation methods of [84, 175] may be applied to solve this problem. Alternatively, instead of estimating the speech signal in a blind fashion, we can exploit the knowledge that one of the signals is speech and that the other is noise, and exploit their different characteristics. In this case, with two microphones, ANC [99, 221] is a well known technique for noise reduction.

As seen in Fig. 9, using a noise reference, the filter $\mathbf{w}$ is adapted to minimize the output power. In practice, a signal-free noise reference is rarely available, resulting in signal leakage in the noise reference path as noted

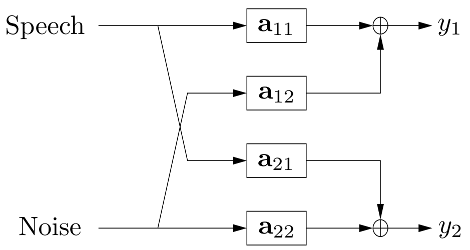

Figure 8: Schematic diagram of the linear convolutive mixing process. $\mathbf{a}_{ij}$ denotes the filter from the j^{th} source to the i^{th} sensor. The noisy signals y_1 and y_2 are observed at the microphones.

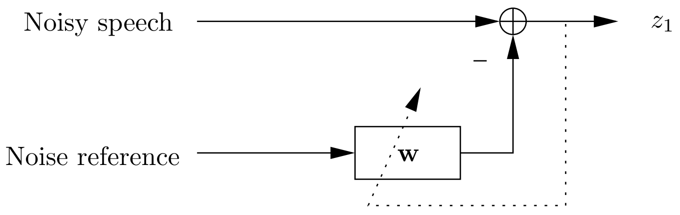

Figure 9: Adaptive noise cancellation.

earlier in the context of the GSC. This causes cancellation of the desired signal at the output. Different approaches have been proposed to solve this problem, e.g., in [164, 234], a second adaptive filter is added to remove the crosstalk in the noise reference. Other methods adapt the filter during time segments when only the noise is present (e.g., during speech pauses detected using a VAD) [1, 47]. One improvement is to perform the adaptation in the frequency domain using bin-wise minimum tracking of any of the diagonal entries of the cross-spectral density of the microphone signals, which we adopt in paper D. A minimum corresponds to speech absence in that particular frequency bin. This approach relies on the observation that speech energy is not present in all frequency bins at all times [144]. The bin-wise minimum tracking is more flexible since it only requires speech to be absent in a particular bin as opposed to VAD based adaptation that requires a noise-only time segment, i.e., speech needs to be absent in all frequency bins simultaneously.

A problem with ANC based methods is that the output is a filtered version of the original speech signal. Using (65) with $S_1(\omega) = S(\omega)$ and $S_2(\omega) = N(\omega)$, the noisy speech input to the ANC can be written as $Y_1(\omega) = a_{11}(\omega)S(\omega) + a_{12}(\omega)N(\omega)$, and the noise reference is $a_{22}(\omega)N(\omega)$. In the ideal case (no speech leakage), after some straightforward calculations, the output of the ANC can be shown to be [204]:

$$Z_1(\omega) = \frac{a_{11}(\omega)a_{22}(\omega) - a_{12}(\omega)a_{21}(\omega)}{a_{22}(\omega) - a_{12}(\omega)} S(\omega), \tag{66}$$

This corresponds to a filtered version of the original speech signal. Instead of this arbitrary filtering, it is desirable to obtain an estimate $a_{11}(\omega)S(\omega)$ which corresponds to the clean speech signal as observed at the microphone.

We note that in the ANC based approach, it is sufficient to estimate only the first row of the unmixing matrix $W(\omega)$ since we are interested in recovering only the speech signal. However, if we estimate the entire unmixing matrix, then it is possible to apply the minimal distortion principle (MDP) [154,155] to obtain a new unmixing matrix $W^{\text{opt}}(\omega) = \text{diag}(W^{-1}(\omega))W(\omega)$. Applying $W^{\text{opt}}(\omega)$ to $\mathbf{Y}(\omega)$ results in $a_{11}(\omega)S(\omega)$ as the output at the first channel. This approach is adopted in paper D, where the unmixing matrix is estimated through bin-wise minimum and maximum tracking using the cross-spectral density of the sensor signals. The resulting approach is a combination of BSS and ANC principles. Through explicit estimation of the speech signal at the first channel by optimizing an energy criterion, the permutation and scaling problems of narrowband BSS are avoided. By estimating the entire unmixing matrix and applying the MDP, the filtering problem of ANC is avoided.

5 Summary of contributions

This thesis deals with the enhancement of speech signals that have been subject to acoustic background noise. An estimation-theoretic approach to exploit prior knowledge about the speech and noise signals is developed using maximum-likelihood and Bayesian MMSE estimation. The use of prior information is shown to result in good performance in practical environments with nonstationary background noise. Both single and multi-microphone speech enhancement techniques are developed. An application of blind source separation concepts to noise reduction is also presented.

Short summaries of the four papers that constitute the main body of the thesis are presented below. All experiments and most of the derivations described in the following papers were performed by the author of this thesis.

Paper A: Codebook driven short-term predictor parameter estimation for speech enhancement

This paper presents an ML approach for the estimation of the speech and noise short-term LP parameters from noisy data and their subsequent use in waveform enhancement schemes. The method exploits a-priori information about speech and noise spectral shapes stored in trained codebooks, parameterized as LP coefficients. The algorithm operates on a frame-by-frame basis, and for each frame, the prior information modelled by the noise codebook is augmented with a long-term estimate of the vector of noise LP coefficients estimated from the noisy observation. This serves as a safety-net for noise types not represented in the codebook. As in [125], prior information is captured only for the spectral shape; the speech and excitation variances are computed on-line. ML estimates of the speech and noise short-term predictor parameters are obtained by searching for the combination of codebook entries that optimizes the likelihood. The estimation involves the computation of the excitation variances of the speech and noise AR models on a frame-by-frame basis, using the a-priori information and the noisy observation. The high computational complexity resulting from a full search of the joint speech and noise codebooks is avoided through an iterative optimization procedure. We introduce a classified noise codebook scheme where different noise codebooks are trained on different noise types, and an appropriate codebook is selected for each frame. Experimental results show that the use of a-priori information and the calculation of the instantaneous speech and noise excitation variances on a frame-by-frame basis result in good performance in both stationary and nonstationary noise conditions.

Paper B: Codebook-based Bayesian speech enhancement for nonstationary environments

In this paper, we propose a Bayesian MMSE approach for the estimation of the short-term predictor parameters of speech and noise, from the noisy observation. We use trained codebooks of speech and noise LP parameters to model the a-priori information required by the Bayesian scheme. In contrast to current Bayesian estimation approaches that consider the excitation variances as part of the a-priori information, in the proposed method they are computed on-line, based on the observation at hand. Consequently, the method performs well in nonstationary noise conditions. The resulting estimates of the speech and noise spectra can be used in a Wiener filter or any state-of-the-art speech enhancement system. We develop both memoryless (using information from the current frame alone) and memory-based (using information from the current and previous frames) estimators. MMSE estimation of functions of the short-term predictor parameters is also addressed, in particular one that leads to the MMSE estimate of the clean speech signal. The classified noise codebook scheme introduced in the ML approach of paper A is employed to select an appropriate noise codebook for each frame. The memory-based estimator has a reduced variance compared to the memoryless estimator. Experiments indicate that the resulting memory-based scheme performs significantly better than competing methods.

Paper C: Multi-channel parametric speech enhancement

A parametric model-based multi-channel approach for speech enhancement is presented. This paper is a generalization of the Bayesian MMSE approach presented in paper B to the multi-channel case. Using multiple microphones allows us to perform both spatial and temporal filtering. The multi-channel Wiener filter can be factorized into an MVDR beamformer followed by a single-channel post-filter. Thus, the estimation is performed using the beamformer output. The MVDR beamformer assumes that the microphone signals are time-aligned (steered) prior to the beamforming. However in practice, ideal alignment is difficult to achieve, and there are steering errors. Therefore we use a signal model that accounts for the effect of steering errors. The model also accounts for a diffuse noise field. By employing an AR model for the speech signal, and using a trained codebook of speech LP coefficients, an MMSE estimate of the clean speech signal is obtained. Robust performance is observed even in the presence of steering errors. Experiments show that the proposed method results in significant performance gains compared to a state-of-the-art diffuse noise post-filter.

Paper D: Speech denoising through source separation and min-max tracking

This paper presents a frequency domain multi-channel noise reduction algorithm based on blind source separation. By tracking the minimum and maximum of the spectral density of the microphone signals in each frequency bin, noise dominated and speech dominated components are identified. The coefficients of the unmixing matrix that are necessary to recover the speech (or noise) are identified from the noise (or speech) dominated components through the optimization of an appropriate energy criterion. The arbitrary filtering of convolutive BSS is compensated using the minimal distortion principle [155]. While it is sufficient to estimate only those unmixing parameters that are required to recover the speech signal, we also estimate the unmixing parameters of the noise signal to be able to apply the minimal distortion principle. Since the proposed method explicitly estimates the speech signal from the noisy mixture, it does not suffer from the permutation problem that is typical to conventional BSS techniques. Experimental results show superior performance compared to a general BSS algorithm.

References

[1] M. J. Al-Kindi and J. Dunlop, "Improved adaptive noise cancellation in the presence of signal leakage on the noise reference channel," *Signal Processing*, vol. 17, no. 3, pp. 241–250, July 1989.

[2] B. D. O. Anderson and J. B. Moore, *Optimal filtering.* Englewood Cliffs, NJ: Prentice Hall, 1979.

[3] D. V. Anderson, "Speech analysis and coding using a multi-resolution sinusoidal transform," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, May 1996, pp. 1037–1040.

[4] D. V. Anderson and M. A. Clements, "Audio signal noise reduction using multi-resolution sinusoidal modeling," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Mar. 1999, pp. 805–808.

[5] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for on-line non-linear/non-gaussian Bayesian tracking," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[6] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, no. 2, pp. 637–655, Aug. 1971.

[7] H. Attias, L. Deng, A. Acero, and J. C. Platt, "A new method for speech denoising and robust speech recognition using probabilistic models for clean speech and for noise," in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, 2001, pp. 1903–1906.

[8] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 13, 2000, pp. 758–764.

[9] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[10] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 4, 1979, pp. 208–211.

[11] J. Bitzer, K. D. Kammeyer, and K. U. Simmer, "An alternative implementation of the superdirective beamformer," in *Proc. IEEE Workshop Appln. Sig. Proc. Audio and Acoustics*, Oct. 1999, pp. 7–10.

[12] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone arrays: signal processing techniques and applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin: Springer-Verlag, 2001, ch. 2, pp. 19–38.

[13] J. Bitzer, K. U. Simmer, and K. D. Kammeyer, "Multi-microphone noise reduction by post-filter and superdirective beamformer," in *Proc. Int. Workshop Acoustic Echo and Noise Control*, Sept. 1999, pp. 100–103.

[14] ——, "Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 5, Mar. 1999, pp. 2965–2968.

[15] ——, “Multi-microphone noise reduction techniques as front-end devices for speech recognition,” *Speech Communication*, vol. 34, no. 1-2, pp. 3–12, Apr. 2001.

[16] P. Bodin and L. F. Villemoes, “Spectral subtraction in the time-frequency domain using wavelet packets,” in *IEEE Workshop on Speech Coding for Telecommunications*, Sept. 1997, pp. 47–48.

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

[18] R. L. Bouquin-Jeannes, A. A. Azirani, and G. Faucon, “Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator,” *IEEE Trans. Speech Audio Processing*, vol. 5, no. 5, pp. 484–487, Sept. 1997.

[19] R. L. Bouquin-Jeannes and G. Faucon, “Proposal of a voice activity detector for noise reduction,” *Electronics Lett.*, vol. 30, no. 12, pp. 930–932, June 1994.

[20] ——, “Study of a voice activity detector and its influence on a noise reduction system,” *Speech Communication*, vol. 16, no. 3, pp. 245–254, Apr. 1995.

[21] B. R. Breed and J. Strauss, “A short proof of the equivalence of LCMV and GSC beamforming,” *IEEE Signal Processing Lett.*, vol. 9, no. 6, pp. 168–169, June 2002.

[22] C. Breithaupt and R. Martin, “MMSE estimation of magnitude-squared DFT coefficients with supergaussian priors,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Apr. 2003, pp. 848–851.

[23] H. Buchner, R. Aichner, and W. Kellermann, “Blind source separation for convolutive mixtures,” in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Boston: Kluwer Academic Publishers, Feb. 2004.

[24] ——, “A generalization of blind source separation algorithms for convolutive mixtures based on second order statistics,” *IEEE Trans. Speech Audio Processing*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

[25] K. M. Buckley, “Broad-band beamforming and the generalized sidelobe canceller,” *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 34, no. 5, pp. 1322–1323, Oct. 1986.

[26] K. M. Buckley and L. J. Griffiths, “An adaptive generalized sidelobe canceller with derivative constraints,” *IEEE Trans. Antennas and Propag.*, vol. 34, no. 3, pp. 311–319, Mar. 1986.

[27] D. Burshtein and S. Gannot, “Speech enhancement using a mixture-maximum model,” *IEEE Trans. Speech Audio Processing*, vol. 10, no. 6, pp. 341–351, Sept. 2002.

[28] O. Cappé, “Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor,” *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[29] J.-F. Cardoso, "Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 5, Apr. 1990, pp. 2655–2658.

[30] ——, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Processing Lett.*, vol. 4, no. 4, pp. 112–114, Apr. 1997.

[31] ——, "Blind signal separation: statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[32] S. Chang, Y. Kwon, S. Yang, and I. Kim, "Speech enhancement for non-stationary noise environment by adaptive wavelet packet," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 2002, pp. 561–564.

[33] Y. M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Apr. 1991, pp. 961–964.

[34] ——, "Speech enhancement based conceptually on auditory evidence," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 39, no. 9, pp. 1943–1954, Sept. 1991.

[35] Y. D. Cho, K. Al-Naimi, and A. Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, May 2001, pp. 737–740.

[36] ——, "Mixed decision-based noise adaptation for speech enhancement," *Electronics Lett.*, vol. 37, no. 8, pp. 540–542, Apr. 2001.

[37] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Lett.*, vol. 8, no. 10, pp. 276–278, Oct. 2001.

[38] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Lett.*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

[39] ——, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[40] ——, "Multichannel post-filtering in nonstationary noise environments," *IEEE Trans. Signal Processing*, vol. 52, no. 5, pp. 1149–1160, May 2004.

[41] ——, "Speech enhancement using a noncausal a priori SNR estimator," *IEEE Signal Processing Lett.*, vol. 11, no. 9, pp. 725–728, Sept. 2004.

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

[43] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.

[44] ——, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Lett.*, vol. 9, no. 2, pp. 12–15, Jan. 2002.

[45] R. R. Coifman and M. V. Wickerhauser, “Entropy-based algorithms for best basis selection,” *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 713–718, Mar. 1992.

[46] P. Comon, “Independent component analysis, a new concept?” *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[47] D. V. Compernolle, “Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Apr. 1990, pp. 833–836.

[48] D. V. Compernolle and S. V. Gerven, “Signal separation in a symmetric adaptive noise canceler by output decorrelation,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 4, Mar. 1992, pp. 221–224.

[49] H. Cox, R. M. Zeskind, and M. M. Owen, “Robust adaptive beamforming,” *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[50] M. Dendrinos, S. Bakamidis, and G. Carayannis, “Speech enhancement from noise: a regenerative approach,” *Speech Communication*, vol. 10, no. 2, pp. 45–57, Feb. 1991.

[51] L. Deng, J. Droppo, and A. Acero, “Log-domain speech feature enhancement using sequential MAP noise estimation and a phase-sensitive model of the acoustic environment,” in *Proc. Int. Conf. on Spoken Language Processing*, Sept. 2002, pp. 1813–1816.

[52] ——, “Recursive noise estimation using iterative stochastic approximation for stereo-based robust speech recognition,” in *Proc. IEEE Workshop Automatic Speech Recognition Understanding*, Dec. 2002, pp. 81–84.

[53] ——, “Incremental Bayes learning with prior evolution for tracking nonstationary noise statistics from noisy speech data,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Apr. 2003, pp. 672–675.

[54] ——, “Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition,” *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 568–580, Nov. 2003.

[55] G. Doblinger, “Computationally efficient speech enhancement by spectral minima tracking in subbands,” in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, vol. 2, Sept. 1995, pp. 1513–1516.

[56] S. Doclo, “Multi-microphone noise reduction and dereverberation techniques for speech applications,” Ph.D. dissertation, Katholieke Universiteit Leuven, Belgium, May 2003.

[57] S. Doclo and M. Moonen, “Robust time-delay estimation in highly adverse acoustic environments,” in *Proc. IEEE Workshop Appln. Sig. Proc. Audio and Acoustics*, Oct. 2001, pp. 59–62.

[58] ——, “Design of far-field and near-field broadband beamformers using eigenfilters,” *Signal Processing*, vol. 83, no. 12, pp. 2641–2673, Dec. 2003.

[59] D. L. Donoho, “De-noising by soft-thresholding,” *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[60] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[61] M. Dorbecker, "Small microphone arrays with optimized directivity for speech enhancement," in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, Sept. 1997, pp. 327–330.

[62] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, pp. 197–208, 2000.

[63] S. C. Douglas, "Blind separation of acoustic signals," in *Microphone arrays: signal processing techniques and applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin: Springer-Verlag, 2001, ch. 16, pp. 355–380.

[64] G. W. Elko, "Microphone array systems for hands-free telecommunication," *Speech Communication*, vol. 20, no. 3-4, pp. 229–240, Dec. 1996.

[65] Y. Ephraim, "A minimum mean square error approach for speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Apr. 1990, pp. 829–832.

[66] ——, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 40, no. 4, pp. 725–735, Apr. 1992.

[67] ——, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

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

[69] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[70] Y. Ephraim and H. L. van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 251–266, July 1995.

[71] S. Fischer and K. D. Kammeyer, "Broadband beamforming with adaptive postfiltering for speech acquisition in noisy environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Apr. 1997, pp. 359–362.

[72] S. Fischer and K. U. Simmer, "Beamforming microphone arrays for speech acquisition in noisy environments," *Speech Communication*, vol. 20, no. 3-4, pp. 215–227, Dec. 1996.

[73] W. Fong, S. Godsill, A. Doucet, and M. West, "Monte Carlo smoothing with application to audio signal enhancement," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 438–449, Feb. 2002.

[74] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, "The voice activity detector for the pan-European digital cellular mobile telephone service," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 1989, pp. 369–372.

[75] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[76] M. Gabrea, "Robust adaptive Kalman filtering-based speech enhancement algorithm," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 2004, pp. 301–304.

[77] M. Gabrea, E. Grivel, and M. Najun, "A single microphone Kalman filter-based noise canceller," *IEEE Signal Processing Lett.*, vol. 6, no. 3, pp. 55–57, Mar. 1999.

[78] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative-batch and sequential algorithms for single microphone speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Apr. 1997, pp. 1215–1218.

[79] ——, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 4, pp. 373–385, July 1998.

[80] ——, "Signal enhancement using beamforming and nonstationarity with application to speech," *IEEE Trans. Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[81] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 6, pp. 561–571, Nov. 2004.

[82] S. Gazor, "Employing Laplacian-Gaussian densities for speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 2004, pp. 297–300.

[83] S. Gazor and W. Zhang, "Speech enhancement employing Laplacian Gaussian mixture," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 896–904, Sept. 2005.

[84] S. V. Gerven and D. V. Compernolle, "Signal separation by symmetric adaptive decorrelation: stability, convergence, and uniqueness," *IEEE Trans. Signal Processing*, vol. 43, no. 7, pp. 1602–1612, July 1995.

[85] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Processing*, vol. 39, no. 8, pp. 1732–1742, Aug. 1991.

[86] E. N. Gilbert and S. P. Morgan, "Optimum design of directive antenna arrays subject to random variations," *Bell Syst. Tech. J.*, pp. 637–663, May 1955.

[87] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Hands free continuous speech recognition in noisy environment using a four microphone array," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 1995, pp. 860–863.

[88] S. Godsill and T. Clapp, "Improvement strategies for Monte Carlo particle filters," in *Sequential Monte Carlo Methods in Practice*, A. Doucet, J. F. G. D. Freitas, and N. J. Gordon, Eds. New York: Springer-Verlag, 2001.

[89] Z. Goh, K.-C. Tan, and T. G. Tan, “Postprocessing method for suppressing musical noise generated by spectral subtraction,” *IEEE Trans. Speech Audio Processing*, vol. 6, no. 3, pp. 287–292, May 1998.

[90] Z. Goh, K.-C. Tan, and B. T. G. Tan, “Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model,” *IEEE Trans. Speech Audio Processing*, vol. 7, no. 5, pp. 510–524, Sept. 1999.

[91] M. Graciarena and H. Franco, “Unsupervised noise model estimation for model-based robust speech recognition,” in *Proc. IEEE Workshop Automatic Speech Recognition Understanding*, Dec. 2003, pp. 351–353.

[92] V. Grancharov, J. Samuelsson, and W. B. Kleijn, “On causal algorithms for speech enhancement,” *IEEE Trans. Speech Audio Processing*, accepted for publication 2005.

[93] Y. Grenier, “A microphone array for car environments,” *Speech Communication*, vol. 12, no. 1, pp. 25–39, Mar. 1993.

[94] L. Griffiths and C. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Trans. Antennas and Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[95] S. Gustafsson, P. Jax, and P. Vary, “A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 1998, pp. 397–400.

[96] S. Gustafsson, R. Martin, P. Jax, and P. Vary, “A psychoacoustic approach to combined acoustic echo cancellation and noise reduction,” *IEEE Trans. Speech Audio Processing*, vol. 10, no. 5, pp. 245–256, July 2002.

[97] P. Handel, “Low-distortion spectral subtraction for speech enhancement,” in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, Sept. 1995, pp. 1549–1552.

[98] J. Hansen and M. Clements, “Constrained iterative speech enhancement with application to automatic speech recognition,” *IEEE Trans. Signal Processing*, vol. 39, no. 4, pp. 795–805, Apr. 1991.

[99] W. A. Harrison, J. S. Lim, and E. Singer, “A new application of adaptive noise cancellation,” *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 34, no. 1, pp. 21–27, Feb. 1986.

[100] R. C. Hendriks, R. Heusdens, and J. Jensen, “Adaptive time segmentation of noisy speech for improved speech enhancement,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Mar. 2005, pp. 153–156.

[101] ——, “Improved subspace based speech enhancement using an adaptive time segmentation,” in *Proc. IEEE First BENELUX/DSP Valley Signal Processing Symposium*, Apr. 2005, pp. 163–166.

[102] W. Herbordt and W. Kellermann, “Frequency-domain integration of acoustic echo cancellation and a generalized sidelobe canceller with improved robustness,” *European Trans. on Telecommunications (ETT)*, vol. 13, no. 2, pp. 123–132, Mar. 2002.

[103] K. Hermus and P. Wambacq, "Assessment of signal subspace based speech enhancement for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 2004, pp. 945–948.

[104] O. Hoshuyama and A. Sugiyama, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, May 1996, pp. 925–928.

[105] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

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

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

[108] J. Huang and Y. Zhao, "Energy-constrained signal subspace method for speech enhancement and recognition," *IEEE Signal Processing Lett.*, vol. 4, no. 10, pp. 283–285, Oct. 1997.

[109] ——, "An energy-constrained signal subspace method for speech enhancement and recognition in white and colored noises," *Speech Communication*, vol. 26, no. 3, pp. 165–181, Nov. 1998.

[110] ——, "A DCT-based fast signal subspace technique for robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 6, pp. 747–751, Nov. 2000.

[111] F. Itakura and S. Saito, "A statistical estimation method for speech spectral density and formant frequencies," *Electron. Comm. Japan*, vol. 53-A, pp. 36–43, 1970.

[112] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 700–708, Nov. 2003.

[113] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 7, pp. 731–740, Oct. 2001.

[114] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 6, pp. 439–448, Nov. 1995.

[115] S. H. Jensen, J. P. Kargo, C. A. Rodbro, and K. V. Sorensen, "Subspace-based speech enhancement with rank-deficient prewhitening," in *Proc. IEEE Speech Coding Workshop*, Oct. 2002, pp. 166–168.

[116] J. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communication*, vol. 6, pp. 314–323, Feb. 1988.

[117] I. M. Johnstone and B. W. Silverman, "Wavelet threshold estimators for data with correlated noise," *J. Royal Statistical Soc. B*, vol. 59, no. 2, pp. 319–351, 1997.

[118] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear estimation.* Prentice Hall, 2000.

[119] M. Kajala and M. Hamalainen, "Broadband beamforming optimization for speech enhancement in noisy environments," in *Proc. IEEE Workshop Appln. Sig. Proc. Audio and Acoustics*, Oct. 1999, pp. 19–22.

[120] J. M. Kates, "An evaluation of hearing-aid array processing," in *Proc. IEEE Workshop Appln. Sig. Proc. Audio and Acoustics*, Oct. 1995, pp. 15–18.

[121] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory.* Prentice Hall, 1993.

[122] J. U. Kim, S. G. Kim, and C. D. Yoo, "The incorporation of masking threshold to subspace speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Apr. 2003, pp. 76–79.

[123] M. Klein and P. Kabal, "Signal subspace speech enhancement with perceptual post-filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 2002, pp. 537–540.

[124] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[125] M. Kuropatwinski and W. B. Kleijn, "Estimation of the excitation variances of speech and noise AR-models for enhanced speech coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 2001, pp. 669–672.

[126] ——, "Minimum mean square error estimation of speech short-term predictor parameters under noisy conditions," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Apr. 2003, pp. 96–99.

[127] W. M. Kushner, V. Goncharoff, C. Wu, V. Nguyen, and J. N. Damoulakis, "The effects of subtractive-type speech enhancement/noise reduction algorithms on parameter estimation for improved recognition and coding in high noise environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 1989, pp. 211–214.

[128] A. Lallouani, M. Gabrea, and C. S. Gargour, "Wavelet based speech enhancement using two different threshold-based denoising," in *Canadian Conference on Electrical and Computer Engineering*, May 2004, pp. 315–318.

[129] K. Y. Lee and S. Jung, "Time-domain approach using multiple Kalman filters and EM algorithm to speech enhancement with nonstationary noise," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 282–291, May 2000.

[130] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Processing Lett.*, vol. 10, no. 4, pp. 104–106, Apr. 2003.

[131] J. S. Lim and A. V. Oppenheim, “All-pole modelling of degraded speech,” *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 26, no. 3, pp. 197–210, June 1978.

[132] ——, “Enhancement and bandwidth compression of noisy speech,” *Proc. IEEE*, vol. 67, no. 12, pp. 7–24, Dec. 1979.

[133] E. Lindemann, “Two microphone nonlinear frequency domain beamformer for hearing aid noise reduction,” in *Proc. IEEE Workshop Appln. Sig. Proc. Audio and Acoustics*, Oct. 1995, pp. 24–27.

[134] P. Lockwood and J. Boudy, “Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars,” *Speech Communication*, vol. 11, no. 2-3, pp. 215–228, June 1992.

[135] B. Logan and T. Robinson, “Adaptive model-based speech enhancement,” *Speech Communication*, vol. 34, no. 4, pp. 351–368, July 2001.

[136] P. Loizou, “Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum,” *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 857–869, Sept. 2005.

[137] T. Lotter and P. Vary, “Noise reduction by maximum a posteriori spectral amplitude estimation with supergaussian speech modeling,” in *Proc. Int. Workshop Acoustic Echo and Noise Control*, Sept. 2003, pp. 83–86.

[138] J. Makhoul, “Spectral analysis of speech by linear prediction,” *IEEE Trans. Audio and Electroacousitcs*, vol. 21, no. 2, pp. 140–148, June 1973.

[139] D. Malah, R. V. Cox, and A. J. Accardi, “Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Mar. 1999, pp. 789–792.

[140] S. Mallat, G. Papicolaou, and Z. Zhang, “Adaptive covariance estimation of locally stationary processes,” *Annals of Statistics*, vol. 26, no. 1, pp. 1–47, Feb. 1998.

[141] C. Marro, Y. Mahieux, and K. U. Simmer, “Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering,” *IEEE Trans. Speech Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.

[142] R. Martin, “An efficient algorithm to estimate the instantaneous SNR of speech signals,” in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, Sept. 1993, pp. 1093–1096.

[143] ——, “Design and optimization of a two microphone speech enhancement system,” in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, Sept. 1995, pp. 2009–2012.

[144] ——, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Trans. Speech Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[145] ——, “Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 2002, pp. 253–256.

[146] ——, “Speech enhancement based on minimum mean-square error estimation and supergaussian priors,” *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 845–856, Sept. 2005.

[147] R. Martin and C. Breithaupt, “Speech enhancement in the DFT domain using Laplacian speech priors.” in *Proc. Int. Workshop Acoustic Echo and Noise Control*, Sept. 2003, pp. 87–90.

[148] R. Martin and R. V. Cox, “New speech enhancement techniques for low bit rate speech coding,” in *Proc. IEEE Speech Coding Workshop*, June 1999, pp. 165–167.

[149] R. Martin, H.-G. Kang, and R. V. Cox, “Low delay analysis/synthesis schemes for joint speech enhancement and low bit rate speech coding,” in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, Sept. 1999, pp. 1463–1466.

[150] R. Martin, D. Malah, R. V. Cox, and A. J. Accardi, “A noise reduction preprocessor for mobile voice communication,” *EURASIP Journal on Applied Signal Processing*, no. 8, pp. 1046–1058, July 2004.

[151] R. Martin, A. Petrovsky, and T. Lotter, “Planar superdirective microphone arrays for speech acquisition in the car,” in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, Sept. 2001, pp. 2623–2626.

[152] R. Martin and P. Vary, “A symmetric two microphone speech enhancement system: theoretical limits and application in a car environment,” in *IEEE Dig. Sig. Proc. Workshop*, Sept. 1992, pp. 4.5.1–4.5.2.

[153] R. Martin, I. Wittke, and P. Jax, “Optimized estimation of spectral parameters for the coding of noisy speech,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 3, June 2000, pp. 1479–1482.

[154] K. Matsuoka, “Independent component analysis and its applications to sound signal separation,” in *Proc. Int. Workshop Acoustic Echo and Noise Control*, Sept. 2003, pp. 15–18.

[155] K. Matsuoka and S. Nakashima, “Minimal distortion principle for blind source separation,” in *Proc. Int. Conf. ICA and BSS*, Dec. 2001, pp. 722–727.

[156] R. J. McAulay and M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[157] R. J. McAulay and T. F. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[158] ——, “Sinusoidal coding,” in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, Amsterdam, 1995, pp. 121–173.

[159] I. A. McCowan and H. Bourlard, “Microphone array post-filter based on noise field coherence,” *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[160] I. A. McCowan, C. Marro, and L. Mauuary, "Robust speech recognition using near-field superdirective beamforming with post-filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 3, June 2000, pp. 1723–1726.

[161] I. A. McCowan and S. Sridharan, "Microphone array sub-band speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 2001, pp. 185–188.

[162] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Apr. 1997, pp. 1167–1170.

[163] D. Middleton and R. Esposito, "Simultaneous optimum detection and estimation of signals in noise," *IEEE Trans. Inform. Theory*, vol. 14, no. 3, pp. 434–444, May 1968.

[164] G. Mirchandani, R. Zinser, and J. Evans, "A new adaptive noise cancellation in the presence of crosstalk," *IEEE Trans. Circuits and Systems II*, vol. 39, no. 10, pp. 681–694, Oct. 1992.

[165] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 2, pp. 159–167, Mar. 2000.

[166] R. A. Monzingo and T. W. Miller, *Introduction to adaptive arrays.* New York: John Wiley and Sons, 1980, ch. 3, pp. 78–154.

[167] D. C. Moore and I. A. McCowan, "Microphone array speech recognition: experiments on overlapping speech in meetings," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 5, Apr. 2003, pp. 497–500.

[168] A. Nadas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 37, no. 10, pp. 1495–1503, Oct. 1989.

[169] T. Nishiura, M. Nakayama, and S. Nakamura, "An evaluation of adaptive beamformer based on average speech spectrum for noisy speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Apr. 2003, pp. 668–671.

[170] S. Nordebo, S. Nordholm, B. Bengtsson, and I. Claesson, "Noise reduction using an adaptive microphone array in a car - a speech recognition evaluation," in *Proc. IEEE Workshop Appln. Sig. Proc. Audio and Acoustics*, Oct. 1993, pp. 16–18.

[171] S. Nordholm, I. Claesson, and B. Bengtsson, "Adaptive array noise suppression of handsfree speaker input in cars," *IEEE Trans. Vehicular Technology*, vol. 42, no. 4, pp. 514–518, Nov. 1993.

[172] S. Nordholm, I. Claesson, and P. Eriksson, "The broad-band Wiener solution for Griffiths-Jim beamformers," *IEEE Trans. Signal Processing*, vol. 40, no. 2, pp. 474–478, Feb. 1992.

[173] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 12, Apr. 1987, pp. 177–180.

[174] K. K. Paliwal and W. B. Kleijn, "Quantization of LPC parameters," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam: Elsevier Science B.V., 1995, ch. 12, pp. 433–468.

[175] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[176] B. L. Pellom and J. H. L. Hansen, "An improved (Auto:I, LSP:T) constrained iterative speech enhancement for colored noise environments," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 6, pp. 573–579, Nov. 1998.

[177] P. M. Peterson, "Using linearly-constrained adaptive beamforming to reduce interference in hearing aids from competing talkers in reverberant rooms," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 12, Apr. 1987, pp. 2364–2367.

[178] D. C. Popescu and I. Zeljkovic, "Kalman filtering of colored noise for speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, May 1998, pp. 997–1000.

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

[180] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, ser. Prentice Hall Signal Processing Series, A. V. Oppenheim, Ed. Upper Saddle River, NJ: Prentice Hall, 2001.

[181] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[182] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, ser. Prentice Hall Signal Processing Series, A. V. Oppenheim, Ed. Englewood Cliffs, NJ: Prentice Hall, 1993.

[183] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 271–287, Apr. 2004.

[184] S. Rangachari, P. C. Loizou, and Y. Hu, "A noise estimation algorithm with rapid adaptation for highly nonstationary environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 2004, pp. 305–308.

[185] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 87–95, Feb. 2001.

[186] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Magazine*, vol. 8, no. 4, pp. 14–38, Oct. 1991.

[187] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 245–257, Apr. 1994.

[188] H. Sameti, H. Sheikhzadeh, and L. Deng, “HMM-based strategies for enhancement of speech signals embedded in nonstationary noise,” *IEEE Trans. Speech Audio Processing*, vol. 6, no. 5, pp. 445–455, Sept. 1998.

[189] M. L. Seltzer, B. Raj, and R. M. Stern, “Likelihood-maximizing beamforming for robust hands-free speech recognition,” *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 489–498, Sept. 2004.

[190] J. W. Seok and K. S. Bae, “Speech enhancement with reduction of noise components in the wavelet domain,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Apr. 1997, pp. 1323–1326.

[191] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, “A parametric formulation of the generalized spectral subtraction method,” *IEEE Trans. Speech Audio Processing*, vol. 6, no. 4, pp. 328–337, July 1998.

[192] K. U. Simmer, J. Bitzer, and C. Marro, “Post-filtering techniques,” in *Microphone arrays: signal processing techniques and applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin: Springer-Verlag, 2001, ch. 3, pp. 39–60.

[193] K. U. Simmer and A. Wasiljeff, “Adaptive microphone arrays for noise suppression in the frequency domain,” in *Second Cost 229 Workshop on Adaptive Algorithms in Communication*, Oct. 1992, pp. 185–194.

[194] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Processing Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[195] J. Sohn and W. Sung, “A voice activity detector employing soft decision based noise spectrum adaption,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 1998, pp. 365–368.

[196] I. Y. Soon, S. N. Koh, and C. K. Yeo, “Noisy speech enhancement using discrete cosine transform,” *Speech Communication*, vol. 24, pp. 249–257, 1998.

[197] P. Sorqvist, P. Handel, and B. Ottersten, “Kalman filtering for low distortion speech enhancement in mobile communication,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Apr. 1997, pp. 1219–1222.

[198] A. Spriet, “Adaptive filtering techniques for noise reduction and acoustic feedback cancellation in hearing aids,” Ph.D. dissertation, K. U. Leuven, Belgium, Sept. 2004.

[199] A. Spriet, M. Moonen, and J. Wouters, “Robustness analysis of multichannel Wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications,” *IEEE Trans. Speech Audio Processing*, vol. 13, no. 4, pp. 487–503, July 2005.

[200] ——, “Stochastic gradient-based implementation of spatially preprocessed speech distortion weighted multichannel Wiener filtering for noise reduction in hearing aids,” *IEEE Trans. Signal Processing*, vol. 53, no. 3, pp. 911–925, Mar. 2005.

[201] T. Sreenivas, “Improved iterative Wiener filtering for non-stationary noise speech enhancement,” in *Proc. Int. Conf. on Spoken Language Processing*, Oct. 2004.

[202] T. Sreenivas and P. Kirnapure, “Codebook constrained Wiener filtering for speech enhancement,” *IEEE Trans. Speech Audio Processing*, vol. 4, no. 5, pp. 383–389, Sept. 1996.

[203] S. Srinivasan and W. B. Kleijn, “Speech enhancement using adaptive time-domain segmentation,” in *Proc. Int. Conf. on Spoken Language Processing*, vol. 2, Oct. 2004, pp. 869–872.

[204] S. Srinivasan, M. Nilsson, and W. B. Kleijn, “Speech denoising through source separation and min-max tracking,” *IEEE Signal Processing Lett.*, submitted.

[205] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, “Codebook driven short-term predictor parameter estimation for speech enhancement,” *IEEE Trans. Speech Audio Processing*, accepted for publication.

[206] ——, “Speech enhancement using a-priori information,” in *Proc. Eurospeech*, Sept. 2003, pp. 1405–1408.

[207] ——, “Estimation of short-term predictor parameters for coding and enhancement of noisy speech,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 2004, pp. 705–708.

[208] ——, “Speech enhancement using a-priori information with classified noise codebooks,” in *Proc. XII European Signal Processing Conf.*, Sept. 2004, pp. 1461–1464.

[209] ——, “Codebook-based Bayesian speech enhancement,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Mar. 2005, pp. 1077–1080.

[210] V. Stahl, A. Fischer, and R. Bippus, “Quantile based noise estimation for spectral subtraction and Wiener filtering,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 3, June 2000, pp. 1875–1878.

[211] M. Sugiyama, “Model based voice decomposition method,” in *Proc. ICSLP*, vol. 4, Oct. 2000, pp. 684–687.

[212] D. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis, “Speech enhancement based on audible noise suppression,” *IEEE Trans. Speech Audio Processing*, vol. 5, no. 6, pp. 497–514, Nov. 1997.

[213] D. Tsoukalas, M. Paraskevas, and J. Mourjopoulos, “Speech enhancement using psychoacoustic criteria,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Apr. 1993, pp. 359–362.

[214] B. D. van Veen and K. M. Buckley, “Beamforming: a versatile approach to spatial filtering,” *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[215] J. Vermaak, C. Andrieu, A. Doucet, and S. Godsill, “Particle methods for Bayesian modelling and enhancement of speech signals,” *IEEE Trans. Speech Audio Processing*, vol. 10, no. 3, pp. 173–185, Mar. 2002.

[216] R. Vetter, “Single channel speech enhancement using MDL-based subspace approach in Bark domain,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 2001, pp. 641–644.

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

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

[219] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 4, pp. 405–413, Oct. 1993.

[220] N. A. Whitmal, J. C. Rutledge, and J. Cohen, "Wavelet-based noise reduction," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 5, May 1995, pp. 3003–3006.

[221] B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, and R. C. Goodlin, "Adaptive noise cancelling: Principles and applications," *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1716, Dec. 1975.

[222] B. Widrow and F.-L. Luo, "Microphone arrays for hearing aids: An overview," *Speech Communication*, vol. 39, no. 1-2, pp. 139–146, Jan. 2003.

[223] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," in *Proc. 11th IEEE SP Workshop on Statistical Signal Processing*, Aug. 2001, pp. 496–499.

[224] J. Yamauchi and T. Shimamura, "Noise estimation using high frequency regions for speech enhancement in low SNR environments," in *Proc. IEEE Speech Coding Workshop*, Oct. 2002, pp. 59–61.

[225] L. Yaroslavsky and Y. Wang, "DFT, DCT, MDCT, DST and signal Fourier spectrum analysis," in *Proc. European Signal Processing Conf.*, Sept. 1998.

[226] C. H. You, S. N. Koh, and S. Rahardja, "Kalman filtering speech enhancement incorporating masking properties for mobile communication in a car environment," in *Proc. IEEE Int. Conf. Multimedia and Expo*, vol. 2, June 2004, pp. 1343–1346.

[227] ——, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 4, pp. 475–486, July 2005.

[228] R. Zelinksi, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 5, Apr. 1988, pp. 2578–2581.

[229] R. Zelinski, "Noise reduction based on microphone array with LMS adaptive post-filtering," *Electronics Lett.*, vol. 26, no. 24, pp. 2036–2037, Nov. 1990.

[230] D. Zhao and W. B. Kleijn, "On noise gain estimation for HMM based speech enhancement," in *Proc. Interspeech 2005 - Eurospeech*, Sept. 2005, pp. 2113–2116.

[231] Y. Zhao, "Maximum likelihood joint estimation of channel and noise for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, June 2000, pp. 1109–1112.

[232] Y. Zhao, S. Wang, and K. C. Yen, "Recursive estimation of time-varying environments for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, May 2001, pp. 225–228.

[233] Y. Zhao, “Frequency-domain maximum likelihood estimation for automatic speech recognition in additive and convolutive noises,” *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 255–266, May 2000.

[234] R. Zinser, G. Mirchandani, and J. Evans, “Some experimental and theoretical results using a new adaptive filter structure for noise cancellation in the presence of crosstalk,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 10, Apr. 1985, pp. 1253–1256.