Model Based Speech Enhancement and Coding

David Yuheng Zhao

赵羽珩

Sound and Image Processing Laboratory
School of Electrical Engineering
KTH (Royal Institute of Technology)

Stockholm 2007
Abstract

In mobile speech communication, adverse conditions, such as noisy acoustic environments and unreliable network connections, may severely degrade the intelligibility and naturalness of the received speech quality, and increase the listening effort. This thesis focuses on countermeasures based on statistical signal processing techniques. The main body of the thesis consists of three research articles, targeting two specific problems: speech enhancement for noise reduction and flexible source coder design for unreliable networks.

Papers A and B consider speech enhancement for noise reduction. New schemes based on an extension to the auto-regressive (AR) hidden Markov model (HMM) for speech and noise are proposed. Stochastic models for speech and noise gains (excitation variance from an AR model) are integrated into the HMM framework in order to improve the modeling of energy variation. The extended model is referred to as a stochastic-gain hidden Markov model (SG-HMM). The speech gain describes the energy variations of the speech phones, typically due to differences in pronunciation and/or different vocalizations of individual speakers. The noise gain improves the tracking of the time-varying energy of non-stationary noise, e.g., due to movement of the noise source. In Paper A, it is assumed that prior knowledge on the noise environment is available, so that a pre-trained noise model is used. In Paper B, the noise model is adaptive and the model parameters are estimated on-line from the noisy observations using a recursive estimation algorithm. Based on the speech and noise models, a novel Bayesian estimator of the clean speech is developed in Paper A, and an estimator of the noise power spectral density (PSD) in Paper B. It is demonstrated that the proposed schemes achieve more accurate models of speech and noise than traditional techniques, and as part of a speech enhancement system provide improved speech quality, particularly for non-stationary noise sources.

In Paper C, a flexible entropy-constrained vector quantization scheme based on Gaussian mixture model (GMM), lattice quantization, and arithmetic coding is proposed. The method allows for changing the average rate in real-time, and facilitates adaptation to the currently available bandwidth of the network. A practical solution to the classical issue of indexing and entropy-coding the quantized code vectors is given. The proposed scheme has a computational complexity that is independent of rate, and quadratic with respect to vector dimension. Hence, the scheme can be applied to the quantization of source vectors in a high dimensional space. The theoretical performance of the scheme is analyzed under a high-rate assumption. It is shown that, at high rate, the scheme approaches the theoretically optimal performance, if the mixture components are located far apart. The practical performance of the scheme is confirmed through simulations on both synthetic and speech-derived source vectors.

Keywords: statistical model, Gaussian mixture model (GMM), hidden Markov model (HMM), noise reduction, speech enhancement, vector quantization.
List of Papers

The thesis is based on the following papers:


In addition to papers A-C, the following papers have also been produced in part by the author of the thesis:


Acknowledgements

Approaching the end of my Ph.D. study, I would like to thank my supervisor, Prof. Bastiaan Kleijn, for introducing me to the world of academic research. Your creativity, dedication, professionalism and hardworking nature have greatly inspired and influenced me. I also appreciate your patience with me, and that you spend many hours of your valuable time for correcting my English mistakes.

I am grateful to all my current and past colleges at the Sound and Image Processing lab. Unfortunately, the list of names is too long to be given here. Thank you for creating a pleasant and stimulating work environment. I particularly enjoyed the cultural diversity and the feeling of an international “family” atmosphere. I am thankful to my co-authors Dr. Jonas Samuelsson, Dr. Mattias Nilsson, Dr. Volodya Grancharov, Jan Plasberg, Dr. Jonas Lindblom, Dr. Alexander Ypma, and Dr. Bert de Vries. Thank you for our fruitful collaborations. Further more, I would like to express my gratitude to Prof. Bastiaan Kleijn, Dr. Mattias Nilsson, Dr. Volodya Grancharov, Tiago Falk and Anders Ekman for proofreading the introduction of the thesis.

During my Ph.D. study, I was involved in projects financially supported by GN ReSound and the Foundation for Strategic Research. I would like to thank Dr. Alexander Ypma, Dr. Bert de Vries, Prof. Mikael Skoglund, Prof. Gunnar Karlsson, for your encouragement and friendly discussions during the projects.

Thanks to my Chinese fellow students and friends at KTH, particularly Xi, Jing, Lei, and Jinfeng. I feel very lucky to get to know you. The fun we have had together is unforgettable.

I would also like to express my deep gratitude to my family, my wife Lili, for your encouragement, understanding and patience; my beloved kids: Andrew and whom yet-to-be-named, for the joy and hope you bring. I thank my parents, for your sacrifice and your endless support; my brother, for all the venture and joy we have together; my mother-in-law, for taking care of Andrew. Last but not least, I am grateful to our host family, Harley and Eva, for your kind hearts and unconditional friendship.

David Zhao, Stockholm, May 2007
Contents

Abstract i
List of Papers iii
Acknowledgements v
Contents vii
Acronyms xi

I Introduction 1
  1 Statistical models .................................. 3
     1.1 Short-term models ................................. 4
     1.2 Long-term models ................................. 6
     1.3 Model-parameter estimation ....................... 9
  2 Speech enhancement for noise reduction ............. 12
     2.1 Overview ......................................... 14
     2.2 Short-term model based approach ................. 17
     2.3 Long-term model based approach ................. 21
  3 Flexible source coding ................................ 25
     3.1 Source coding ..................................... 25
     3.2 High-rate theory .................................. 30
     3.3 High-rate quantizer design ....................... 32
  4 Summary of contributions ............................. 35
References .................................................. 37

II Included papers 51

A HMM-based Gain-Modeling for Enhancement of Speech in Noise A1
  1 Introduction ......................................... A1
## Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3GPP</td>
<td>The Third Generation Partnership Project</td>
</tr>
<tr>
<td>AMR</td>
<td>Adaptive Multi-Rate</td>
</tr>
<tr>
<td>AR</td>
<td>Auto-Regressive</td>
</tr>
<tr>
<td>AR-HMM</td>
<td>Auto-Regressive Hidden Markov Model</td>
</tr>
<tr>
<td>CCR</td>
<td>Comparison Category Rating</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative Distribution Function</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier Transform</td>
</tr>
<tr>
<td>ECVQ</td>
<td>Entropy-Constrained Vector Quantizer</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation Maximization</td>
</tr>
<tr>
<td>EVRC</td>
<td>Enhanced Variable Rate Codec</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>IEEE</td>
<td>Institute of Electrical and Electronics Engineers</td>
</tr>
<tr>
<td>I.I.D.</td>
<td>Independent and Identically-Distributed</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>KLT</td>
<td>Karhunen-Loève Transform</td>
</tr>
<tr>
<td>LL</td>
<td>Log-Likelihood</td>
</tr>
<tr>
<td>LP</td>
<td>Linear Prediction</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear Prediction Coefficient</td>
</tr>
<tr>
<td>LSD</td>
<td>Log-Spectral Distortion</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>-------------</td>
<td>------------------------------------------------</td>
</tr>
<tr>
<td>LSF</td>
<td>Line Spectral Frequency</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum A-Posteriori</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>MMSE</td>
<td>Minimum Mean Square Error</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean Opinion Score</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
</tr>
<tr>
<td>MVU</td>
<td>Minimum Variance Unbiased</td>
</tr>
<tr>
<td>PCM</td>
<td>Pulse-Code Modulation</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability Density Function</td>
</tr>
<tr>
<td>PESQ</td>
<td>Perceptual Evaluation of Speech Quality</td>
</tr>
<tr>
<td>PMF</td>
<td>Probability Mass Function</td>
</tr>
<tr>
<td>PSD</td>
<td>Power Spectral Density</td>
</tr>
<tr>
<td>RCVQ</td>
<td>Resolution-Constrained Vector Quantizer</td>
</tr>
<tr>
<td>REM</td>
<td>Recursive Expectation Maximization</td>
</tr>
<tr>
<td>SG-HMM</td>
<td>Stochastic-Gain Hidden Markov Model</td>
</tr>
<tr>
<td>SMV</td>
<td>Selectable Mode Vocoder</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>SSNR</td>
<td>Segmental Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>SS</td>
<td>Spectral Subtraction</td>
</tr>
<tr>
<td>STSA</td>
<td>Short-Time Spectral Amplitude</td>
</tr>
<tr>
<td>VAD</td>
<td>Voice Activity Detector</td>
</tr>
<tr>
<td>VoIP</td>
<td>Voice over IP</td>
</tr>
<tr>
<td>VQ</td>
<td>Vector Quantizer</td>
</tr>
<tr>
<td>WF</td>
<td>Wiener Filtering</td>
</tr>
<tr>
<td>WGN</td>
<td>White Gaussian Noise</td>
</tr>
</tbody>
</table>
Part I

Introduction
Introduction

The advance of modern information technology is revolutionizing the way we communicate. Global deployment of mobile communication systems has enabled speech communication from and to nearly every corner of the world. Speech communication systems based on the Internet protocol (IP), the so-called voice over IP (VoIP), are rapidly emerging. The new technologies allow for communication over the wide range of transport media and protocols that is associated with the increasingly heterogeneous communication network environment. In such a complex communication infrastructure, adverse conditions, such as noisy acoustic environments and unreliable network connections, may severely degrade the intelligibility and naturalness of the perceived speech quality, and increase the required listening effort for speech communication. The demand for better quality in such situations has given rise to new problems and challenges. In this thesis, two specific problems are considered: enhancement of speech in noisy environments and source coder design for unreliable network conditions.

Figure 1 illustrates an adverse speech communication scenario that is

![Figure 1: A one-direction example of speech communication in a noisy environment.](image-url)
typical of mobile communications. The speech signal is recorded in an environment with a high level of ambient noise such as on a street, in a restaurant, or on a train. In such a scenario, the recorded speech signal is contaminated by additive noise and has degraded quality and intelligibility compared to the clean speech. Speech quality is a subjective measure on how pleasant the signal sounds to the listeners. It can be related to the amount of effort required by listeners to understand the speech. Speech intelligibility can be measured objectively based on the percentage of sounds that are correctly understood by listeners. Naturally, both speech quality and intelligibility are important factors in the subjective rating of a speech signal by a listener. The level of background noise may be suppressed by a speech enhancement system. However, suppression of the background noise may lead to reduced speech intelligibility, due to possible speech distortion associated with excessive noise suppression. Therefore, most practical speech enhancement systems for noise reduction are designed to improve the speech quality, while preserving the speech intelligibility [6, 60, 61]. Besides speech communication, speech enhancement for noise reduction may be applied to a wide range of other speech processing applications in adverse situations, including speech recognition, hearing aids and speaker identification.

Another limiting factor for the quality of speech communication is the network condition. It is a particularly relevant issue for wireless communication, as the network capacity may vary depending on interferences from the physical environment. The same issue occurs in heterogeneous networks with a large variety of different communication links. Traditional source coder design is often optimized for a particular network scenario, and may not perform well in another network. To allow communication with the best possible quality in the new network, it is necessary to adapt the coding strategy, e.g., coding rate, depending on the network condition.

This thesis concerns statistical signal processing techniques that facilitate better solutions to the aforementioned problems. Formulated in a statistical framework, speech enhancement is an estimation problem. Source coding can be seen as a constrained optimization problem that optimizes the coder under certain rate constraints. Statistical modeling of speech (and noise) is a key element for solving both problems.

The techniques proposed in this thesis are based on a hidden Markov model (HMM) or a Gaussian mixture model (GMM). Both are generic models capable of modeling complex data sources such as speech. In Papers A and Paper B, we propose an extension to an HMM that allows for more accurate modeling of speech and noise. We show that the improved models also lead to an improved speech enhancement performance. In Paper C, we propose a flexible source coder design for entropy-constrained vector quantization. The proposed coder design is based on Gaussian mixture modeling of the source.

The introduction is organized as follows. In section 1, an overview of
Statistical modeling for speech is given. We discuss short-term models for modeling the local statistics of a speech frame, and long-term models (such as an HMM or a GMM) that describe the long-term statistics spanning over multiple frames. Parameter estimation using the expectation-maximization (EM) algorithm is discussed. In section 2, we introduce speech enhancement techniques for noise reduction, with our focus on statistical model-based methods. Different approaches using short-term and long-term models are reviewed and discussed. Finally, in section 3, an overview of theory and practice for flexible source coder design is given. We focus on the high-rate theory and coders derived from the theory.

1 Statistical models

Statistical modeling of speech is a relevant topic for nearly all speech processing applications, including coding, enhancement, and recognition. A statistical model can be seen as a parameterized set of probability density functions (PDFs). Once the parameters are determined, the model provides a PDF of the speech signal that describes the stochastic behavior of the signal. A PDF is useful for designing optimal quantizers [64,91,128], estimators [35,37–39,90], and recognizers [11,105,108] in a statistical framework.

One can make statistical models of speech in different representations. The simplest one applies directly to the digital speech samples in the waveform domain (analog speech is out of the scope of this thesis). For instance, a histogram of speech waveform samples, as shown in Figure 2, provides
the long-term averaged distribution of the amplitudes of speech samples. A similar figure is shown in [135]. The distribution is useful for designing an entropy code for a pulse-code modulation (PCM) system. However, in such a model, statistical dependencies among the speech samples are not explicitly modeled.

It is more common to model speech signal as a stochastic process that is locally stationary within 20-30 ms frames. The speech signal is then processed on a frame-by-frame basis. Statistical models in speech processing can be roughly divided into two classes, short-term and long-term signal models, depending on their scope of operation. A short-term model describes the statistics of the vector for a particular frame. As the statistics are changing over the frames, the parameters of a short-term model are to be determined for each frame. On the other hand, a model that describes the statistics over multiple signal frames is referred to as a long-term model.

1.1 Short-term models

It is essential to model the sample dependencies within a speech frame of 20-30 ms. This short-term dependency determines the formants of the speech frame, which are essential for recognizing the associated phoneme. Modeling of this dependency is therefore of interest to many speech applications.

Auto-regressive model

An efficient tool for modeling the short-term dependencies in the time domain is the auto-regressive (AR) model. Let $x[i]$ denote the discrete-time speech signal, the AR model of $x[i]$ is defined as

$$x[i] = - \sum_{j=1}^{p} \alpha[j]x[i-j] + e[i], \quad (1)$$

where $\alpha$ are the AR model parameters, also called the linear prediction coefficients, $p$ is the model order and $e[i]$ is the excitation signal, modeled as white Gaussian noise (WGN). An AR process is then modeled as the output of filtering WGN through an all-pole AR model filter.

For a signal frame of $K$ samples, the vector $x = [x[1], \ldots, x[K]]$ can be seen as a linearly transformation of a WGN vector $e = [e[1], \ldots, e[K]]$, by considering the filtering process as a convolution formulated as a matrix transform. Consequently, the PDF of $x$ can be modeled as a multi-variate Gaussian PDF, with zero-mean and a covariance matrix parameterized using the AR model parameters. The AR Gaussian density function is defined as

$$f(x) = \frac{1}{(2\pi \sigma^2)^{\frac{K}{2}}} \exp \left( -\frac{1}{2}x^T D^{-1}x \right), \quad (2)$$
with the covariance matrix $D = \sigma^2(A^2A)^{-1}$, where $A$ is a $K \times K$ lower triangular Toeplitz matrix with the first $p + 1$ elements of the first column consisting of the AR coefficients $[\alpha[0], \alpha[1], \alpha[2], \ldots, \alpha[p]]^T$, $\alpha[0] = 1$, and the remaining elements equal to zero, $\sharp$ denotes the Hermitian transpose, and $\sigma^2$ denotes the excitation variance. Due to the structure of $A$, the covariance matrix $D$ has the determinant $\sigma^{2K}$.

Applying the AR model for speech processing is often motivated and supported by the physical properties of speech production [43]. A commonly used model of speech production is the source-filter model as shown in Figure 3. The excitation signal, modeled using a pulse train for a voiced sound or Gaussian noise for an unvoiced sound, is filtered through a time-varying all-pole filter that models the vocal tract. Due to the absence of a pitch model, the AR Gaussian model is more accurate for unvoiced sounds.

**Frequency-domain models**

The spectral properties of a speech signal can be analyzed in the frequency domain through the short-time discrete Fourier transform (DFT). On a frame-by-frame basis, each speech segment is transformed into the frequency domain by applying a sliding window to the samples and a DFT on the windowed signal.

A commonly used model is the complex Gaussian model, motivated by the asymptotic properties of the DFT coefficients. Assuming that the DFT length approaches infinity and that the span of correlation of the signal frame is short compared to the DFT length, the DFT coefficients may be considered as a weighted sum of weakly dependent random samples [51, 94, 99, 100]. Using the central limit theorem for weakly dependent variables [12, 130], the DFT coefficients across frequency can be considered as independent.
zero-mean Gaussian random variables. For \( k \not\in \{1, \frac{k}{2} + 1\} \), the PDF of the DFT coefficients of the \( k \)th frequency bin \( X[k] \) is given by

\[
f(X[k]) = \frac{1}{\pi \lambda^2[k]} \exp\left(-\frac{|X[k]|^2}{\lambda^2[k]} \right),
\]

where \( \lambda^2[k] = E[|X[k]|^2] \) is the power spectrum. For \( k \in \{1, \frac{k}{2} + 1\} \), \( X[k] \) is real and has a Gaussian distribution with zero mean and variance \( E[|X[k]|^2] \). Due to the conjugate symmetry of DFT for real signals, only half of the DFT coefficients need to be considered.

One consequence of the model is that each frequency bin may be analyzed and processed independently. Although, for typical speech processing DFT lengths, the independence assumption does not always hold, it is widely used, since it significantly simplifies the algorithm design and reduces the associated computational complexity. For instance, the complex Gaussian model has been successfully applied in speech enhancement [38, 39, 92].

The frequency domain model of an AR process can be derived under the asymptotic assumption. Assuming that the frame length approaches infinity, \( A \) is well approximated by a circulant matrix (neglecting the frame boundary), and is approximately diagonalized by the discrete Fourier transform. Therefore, the frequency domain AR model is of the form of (3), and the power spectrum \( \lambda^2[k] \) is given by

\[
\lambda^2[k] = \sigma^2 \left| \sum_{n=0}^{p} a[n] e^{-jn\nu k} \right|^2.
\]

It has been argued [89, 103, 118] that supergaussian density functions, such as the Laplace density and the two-sided Gamma density, fit better to speech DFT coefficients. However, the experimental studies were based on the long-term averaged behavior of speech, and may not necessarily hold for short-time DFT coefficients. Speech enhancement methods based on supergaussian density function have been proposed in [19, 85, 86, 89, 90]. The comprehensive evaluations in [90] suggest that supergaussian methods achieve consistent but small improvement in terms of higher segmental SNR results relative to the Wiener filtering approach (which is based on the Gaussian model). However, in terms of perceptual quality, the Ephraim-Malah short-time spectral amplitude estimators [38, 39] based on the Gaussian model were found to provide more pleasant residual noise in the processed speech [90].

1.2 Long-term models

With a long-term model, we mean a model that describes the long-term statistics that span over multiple signal frames of 20-30 ms. The short-term models in section 1.1 apply to one signal frame only, and the model
parameters obtained for a frame may not generalize to another frame. A long-term model, on the other hand, is generic and flexible enough to allow statistical variations in different signal frames. To model different short-term statistics in different frames, a long-term model comprises multiple states, each containing a sub-model describing the short-term statistics of a frame. Applied to speech modeling, each sub-model represents a particular class of speech frames that are considered statistically similar. Some of the most well-known long-term models for speech processing are the hidden Markov model (HMM), and the Gaussian mixture model (GMM).

Hidden Markov model

The hidden Markov model (HMM) was originally proposed by Baum and Petrie [13] as probabilistic functions of finite state Markov chains. HMM applied for speech processing can be found in automatic speech recognition [11, 65, 105] and speech enhancement [35, 36, 111, 147].

An HMM models the statistics of an observation sequence, \( x_1, \ldots, x_N \). The model consists of a finite number of states that are not observable, and the transition from one state to another is controlled by a first-order Markov chain. The observation variables are statistically dependent of the underlying states, and may be considered as the states observed through a noisy channel. Due to the Markov assumption of the hidden states, the observations are considered to be statistically interdependent to each other. For a realization of state sequence, let \( s = [s_1, \ldots, s_N] \), \( s_n \) denote the state of frame \( n \), \( a_{s_{n-1}s_n} \) denote the transition probability from state \( s_{n-1} \) to \( s_n \) with \( a_{s_0s_1} \) being the initial state probability, and let \( f_{s_n}(x_n) \) denote the output probability of \( x_n \) for a given state \( s_n \), the PDF of the observation sequence \( x_1, \ldots, x_N \) of an HMM is given by

\[
f(x_1, \ldots, x_N) = \sum_{s \in S} \prod_{n=1}^{N} a_{s_{n-1}s_n} f_{s_n}(x_n),
\]

where the summation is over the set of all possible state sequences \( S \). The joint probability of a state sequence \( s \) and the observation sequence \( x_1, \ldots, x_N \) consists of the product of the transition probabilities and the output probabilities. Finally, the PDF of the observation sequence is the sum of the joint probabilities over all possible state sequences.

Applied to speech modeling, each sub-model of a state represents the short-term statistics of a frame. Using the frequency domain Gaussian model (3), each state consists of a power spectral density representing a particular class of speech sounds (with similar spectral properties). The Markov chain of states models the temporal evolution of short-term speech power spectra.
Figure 4: Examples of a one-dimensional GMM (left) and a two-dimensional GMM (right).

Gaussian mixture model

Gaussian mixture modeling is a convenient tool useful in a wide range of applications. An early application of GMM was found in adaptive block quantization applied to image coding [131]. More recent examples of applications include clustering analysis [150], image coding [126], image retrieval [55], multiple description coding [115], speaker identification [108], speech coding [80, 112, 114], speech quality estimation [42], speech recognition [33, 34], and vector quantization [58, 127, 128, 148].

A Gaussian mixture model (GMM) is defined as a weighted sum of Gaussian densities,

\[ f(x) = \sum_{m=1}^{M} \rho_m f_m(x), \]

where \( m \) denotes the component index, \( \rho \) denote the mixture weights, and \( f_m(x) \) are the component Gaussian PDFs. Examples of a one-dimensional GMM and a two-dimensional GMM are shown in Figure 3.

A GMM can be interpreted in two conceptually different ways. It can be seen as a generic PDF that has the ability of approximate an unknown PDF by adjusting the model parameters over a database of observations. Another interpretation of a GMM is as a special case of an HMM, by assuming independent and identically-distributed (I.I.D.) random variables, such that \( f(x_1, \ldots, x_N) = f(x_1) \cdots f(x_N) \). The model can then be formulated as an HMM by assigning the initial state probability according to the mixture component weights, and assuming the matrix of state transition probabilities to be the identity matrix. In fact, to generate random
observations from a GMM, one can first generate an I.I.D. state sequence according to the mixture component weights, and then generate the observation sequence according to the component PDF of each state sample. The generated observations will be distributed according to the GMM. Applied to speech modeling, each component may represent a class of speech sounds, the weights are according to the probability of their appearance, and each component PDF describes the power spectrum of the sound class.

The main difference between a GMM and an HMM with Gaussian submodels is that, in a GMM, consecutive frames are considered independent, while in an HMM, the Markov assumption of the underlying state sequence models their temporal dependency. An HMM is therefore a more general and powerful model than a GMM. Nevertheless, GMMs are popular models in speech processing due to their simplicity in analysis and good modeling capabilities. In fact, GMMs are often used as sub-models of an HMM in practical implementations of HMM based methods.

1.3 Model-parameter estimation

Using an HMM or a GMM, the joint PDF of a sequence $x_1, \ldots, x_N$, denoted $f(x_1, \ldots, x_N)$, is parameterized by unknown parameters, denoted by $\theta$. Each value of $\theta$ corresponds to one particular PDF of $x_1, \ldots, x_N$. Since the true PDF is unknown, the value of $\theta$ is to be determined from data. Let $\hat{\theta}$ denote the estimated parameters, as a function of $x_1, \ldots, x_N$. A commonly used criterion of an optimal estimator is that 1) the estimator is unbiased, i.e.,

$$E[\hat{\theta}] = \theta, \quad (7)$$

and 2) the estimator gives an estimate that has the minimum variance. Such an estimator is referred to as the minimum variance unbiased (MVU) estimator [67]. Unfortunately, the MVU estimator is difficult to determine in many practical applications. An alternative estimator is the maximum likelihood (ML) estimator, which maximizes the likelihood function or equivalently the logarithm of the likelihood function (log-likelihood function),

$$\hat{\theta} = \arg \max_{\theta} f(x_1, \ldots, x_N | \theta) \quad (8)$$

$$= \arg \max_{\theta} \log f(x_1, \ldots, x_N | \theta), \quad (9)$$

where $f(x_1, \ldots, x_N | \theta)^1$ denotes the likelihood function, viewed as a function of $\theta$ for the given data sequence $x_1, \ldots, x_N$. The ML approach typically leads to practical estimators that can be implemented for complex estimation tasks. Also, it can be shown [67] that the ML estimator is asymptoti-

---

1The likelihood function can also be seen as the PDF of $x_1, \ldots, x_N$ conditioned on the parameter vector $\theta$. 
cally optimal, and approaches the performance of the MVU estimator when the data set is large.

For a generic PDF based on GMM or HMM, directly solving (8) or (9) leads to intractable analytical expressions. A more practical and commonly used method is the expectation-maximization (EM) algorithm [31].

**Expectation-maximization (EM) algorithm**

The expectation-maximization (EM) algorithm originated from early work of Dempster et. al [31]. The EM algorithm is used in Papers A and B for parameter estimation of an extended HMM model of speech. The EM algorithm applies to a given batch of data, and is suitable for off-line optimization of the model. For other model parameters, e.g., in a noise model, the observation data may not be available off-line and on-line estimation using a recursive algorithm is necessary. The recursive formulation of the EM algorithm was proposed in [132], and applied to on-line HMM parameter estimation in [72]. The recursive EM algorithm is used in Paper A for estimation of the gain model parameters, and in Paper B for estimation of noise model parameters. Both algorithms are based on the same principle of incomplete data. Herein, an introduction to the theory behind the EM algorithm is given.

The EM algorithm is an iterative algorithm that ensures non-decreasing likelihood scores during each iteration step. Since the likelihood function is upper-bounded, the algorithm guarantees convergence to a locally optimal solution. The EM algorithm is designed for the estimation scenario when the observation sequence is incomplete or partially missing. The missing observation sequence can be either real or artificial. In either case, the EM algorithm is applicable when the ML estimator leads to intractable analytical formulas but is significantly simplified when assuming the existence of additional unknown values.

For notational convenience, let $\mathcal{X}_1^N = \{x_1, \ldots, x_N\}$ denote the observation data, $\mathcal{Z}_1^N = \{z_1, \ldots, z_N\}$ denote the missing observation data. The complete data is then given by $\{\mathcal{X}_1^N, \mathcal{Z}_1^N\}$, and $\log f(\mathcal{X}_1^N, \mathcal{Z}_1^N|\theta)$ is the complete log-likelihood function. Let $\hat{\theta}^{(j)}$ denote the estimated parameter vector from the $j$’th EM iterations. The EM algorithm consists of two sequential sub-steps within each iteration. The E-step (expectation step) formulates the auxiliary function, denoted $Q(\theta|\hat{\theta}^{(j)})$, which is the expected value of the complete data log-likelihood with respect to the missing data given the observation data and the parameter vector estimate of the $j$’th iteration,

$$Q(\theta|\hat{\theta}^{(j)}) = \int_{\mathcal{Z}_1^N} f(\mathcal{Z}_1^N|\mathcal{X}_1^N, \hat{\theta}^{(j)}) \log \left( f(\mathcal{Z}_1^N, \mathcal{X}_1^N|\theta) \right) d\mathcal{Z}_1^N. \quad (10)$$

The maximization step in the EM algorithm gives a new estimate of the
parameter vector by maximizing the $Q(\theta|\hat{\theta}^{(j)})$ function of the current step,

$$\hat{\theta}^{(j+1)} = \arg \max_\theta Q(\theta|\hat{\theta}^{(j)}). \tag{11}$$

It can be noted that the expectation of (10) is evaluated with respect to $f(Z_1^N|X_1^N, \hat{\theta}^{(j)})$ using the $j$’th parameter estimate, and the maximization of (11) is over $\theta$ in the complete log-likelihood function, $\log f(Z_1^N, X_1^N|\theta)$. Therefore, it is essential to select $Z_1^N$ such that differentiating (11) with respect to $\theta$ and setting the resulting expression to zero should lead to tractable solutions for the update equations.

Estimation of both GMM and HMM parameters can be derived using the EM framework. For a GMM, the missing data consists of the component indices from which the observation data is generated. For an HMM, the missing data is the state sequence. For a detailed derivation of the EM algorithm for GMM and HMM, we refer to [17]. The EM algorithm and the recursive EM algorithm are extensively used in Papers A and B to estimate the model parameters of extended HMMs of speech and noise.

**Convergence of the EM algorithm**

The iterative procedure of the EM algorithm ensures convergence towards a locally optimal solution. A proof of convergence is given in [31], and the proof is outlined in this section. We first show that the log-likelihood function (9) over a large data set is upper-bounded. We then show that the EM iterations improve the log-likelihood function in each iteration step.

Assume existence of a data set consisting of a large number of $X_1^N$, where each data sequence $X_1^N$ is distributed according to a “true” distribution function $f(X_1^N)$. Note the difference to $f(X_1^N|\theta)$, which is a PDF as a function of an unknown parameter vector $\theta$. The log-likelihood function of this data set is

$$\mathcal{L} = \log \prod_{X_1^N} f(X_1^N|\theta) = \sum_{X_1^N} \log f(X_1^N|\theta) \approx \int_{X_1^N} f(X_1^N) \log f(X_1^N|\theta) dX_1^N. \tag{12}$$

Comparing the log-likelihood function with the expectation of the logarithmic transform of the true PDF $\int_{X_1^N} f(X_1^N) \log f(X_1^N)$, we get

$$\mathcal{L} - \int_{X_1^N} f(X_1^N) \log f(X_1^N) = \int_{X_1^N} f(X_1^N) \log \frac{f(X_1^N|\theta)}{f(X_1^N)} dX_1^N \leq \int_{X_1^N} f(X_1^N) \left( \frac{f(X_1^N|\theta)}{f(X_1^N)} - 1 \right) dX_1^N = 1 - 1 = 0, \tag{13}$$
where the inequality is due to \( \log x \leq x - 1 \). Hence, the log-likelihood function is upper bounded by the expectation of the logarithmic transform of the true PDF \( \int f(X_i^N) \log f(X_i^N) \).

The log-likelihood score of the \( j \)th iteration can be written as

\[
\log f(X_i^N|\hat{\theta}^{(j)}) = \left( \log f(X_i^N|\hat{\theta}^{(j)}) \right) \int f(Z_i^N|X_i^N, \hat{\theta}^{(j)}) dZ_i^N
\]

where the last step uses \( \log f(X_i^N|\hat{\theta}^{(j)}) \) and the inequality is due to Jensen’s inequality.

Comparing the log-likelihood score of the \( j+1 \)'th iteration to the previous iteration, we get

\[
\log f(X_i^N|\hat{\theta}^{(j+1)}) = \log \left( \int f(X_i^N, Z_i^N|\hat{\theta}^{(j+1)}) dZ_i^N \right)
\]

\[
\geq \int f(Z_i^N|X_i^N, \hat{\theta}^{(j)}) \log \left( \frac{f(X_i^N, Z_i^N|\hat{\theta}^{(j+1)})}{f(Z_i^N|X_i^N, \hat{\theta}^{(j)})} \right) dZ_i^N
\]

where the inequality is due to Jensen’s inequality.

Comparing the log-likelihood score of the \( (j+1) \)'th iteration to the previous iteration, we get

\[
\log f(X_i^N|\hat{\theta}^{(j+1)}) - \log f(X_i^N|\hat{\theta}^{(j)})
\]

\[
\geq \int f(Z_i^N|X_i^N, \hat{\theta}^{(j)}) \log \left( \frac{f(X_i^N, Z_i^N|\hat{\theta}^{(j+1)})}{f(X_i^N, Z_i^N|\hat{\theta}^{(j)})} \right) dZ_i^N
\]

\[
= Q(\hat{\theta}^{(j+1)}|\hat{\theta}^{(j)}) - Q(\hat{\theta}^{(j)}|\hat{\theta}^{(j)}) \geq 0,
\]

where the last step follows from the definition of the maximization step (11). Since the log-likelihood score is upper bounded, the convergence is proven.

2 Speech enhancement for noise reduction

In Papers A and B, we apply the statistical framework for enhancement of speech in noisy environments. Noise reduction has become an increasingly
important component in speech communication systems. Based on a statistical framework using models of speech and noise, a noise reduction system can be formulated as an estimation problem, in which the clean speech is to be estimated from the noisy observation signal. In estimation theory, the performance of an estimator is commonly measured as the expected value of a distortion function that quantifies the difference between the clean speech signal and the estimated signal. Success of the estimator that minimizes the expected distortion depends largely on the accuracy of the assumed speech and noise models and the perceptual relevance of the distortion measure.

Due to the inherent complexity of natural speech and human perception, the present models and distortion measures are only simplified approximations. A large variety of noise reduction methods has been proposed and their differences are often due to different assumptions on the models and distortion measures.

The models discussed in section 1 can be applied to modeling of the statistics of speech and noise. In the design of a noise reduction system, additional assumptions on the noisy signal are made. Depending on the environment, the interfering noise may be additive, e.g., from a separate noise generator, or convolutive, e.g., due to room reverberations or fading. Depending on the number of available microphones, the observation data consist of one or more noisy signals. The corresponding noise reduction methods are divided into single-channel (one microphone) and multi-channel (microphone array) systems. The focus of this thesis is on single-channel noise reduction for an additive noise environment. The noisy speech signal is assumed to be the sum of the speech and noise signals. Furthermore, the additive noise is often considered to be statistically independent of speech.

A successful noise reduction system may be used as a preprocessor to a speech communication system or an automatic speech recognition system. In fact, several algorithms have been successfully integrated into standardized speech coders for mobile speech communication. Examples of speech coders with a noise reduction module include the Enhanced Variable Rate codec (EVRC) [1] and Selectable Mode Vocoder (SMV) [7]. 3GPP (The third Generation Partnership Project) has standardized minimum performance requirements and an evaluation procedure for the Adaptive Multi-Rate (AMR) codec [6], and a number of noise reduction algorithms that satisfy the requirements have been developed, e.g., [3–5, 8].

The remainder of this section is organized as follows. An overview of single-channel noise reduction is given in section 2.1. The focus is on frequency domain approaches that make use of a short-time spectral attenuation filter. The Bayesian estimation framework is introduced and noise estimation is discussed. Sections 2.2 and 2.3 provide an overview of classical statistical model based methods using short-term models and long-term models, respectively. Sections 2.2 and 2.3 also serves as an introduction to Papers A and B. For a more extensive tutorial on speech enhancement
algorithms for noise reduction, we refer to [37, 78, 135].

2.1 Overview
Due to the quasi-stationarity property of speech, the noisy signal is often processed on a frame-by-frame basis. For additive noise, the noisy speech signal of the \( n \)’th frame, is modeled as

\[
y_n[i] = x_n[i] + w_n[i],
\]

(23)

where \( i \) is the time index of the \( n \)’th frame and \( y, x, w \) denote noisy speech, clean speech and noise, respectively. In the vector notation, the noisy speech, clean speech and noise frames are denoted by \( y_n, x_n \) and \( w_n \), respectively. In the following sections, the speech model parameters are denoted using overbar \( \bar{\cdot} \) and the noise model parameters are denoted using double dots \( \ddot{\cdot} \).

Short-time spectral attenuation
It is common to perform estimation in the spectral domain, e.g., through short-time Fourier analysis and synthesis. An analysis window is applied to each time domain signal segment and the DFT is then applied. For the \( n \)’th frame, the frequency domain signal model is given by

\[
Y_n[k] = X_n[k] + W_n[k],
\]

(24)

where \( k \) denotes the index of the frequency band.

For frequency-domain noise reduction, the noisy spectrum is attenuated in the frequency bins that contain noise. The level of attenuation depends on the signal to noise ratio (SNR) of the frequency bin. The attenuation is performed by means of an adaptive spectral attenuation filter, denoted by \( H_n[k] \), that is applied to \( Y_n[k] \) to obtain an estimate of \( X_n[k], \hat{X}_n[k] \), given by:

\[
\hat{X}_n[k] = H_n[k]Y_n[k].
\]

(25)

The inverse DFT is applied to \( \hat{X}_n = [\hat{X}_n[1], \ldots, \hat{X}_n[K]]^T \) to obtain the enhanced speech signal frame \( \hat{x}_n \) in the time domain. To avoid discontinuity at frame boundaries, the overlap and add approach is used in synthesis. The schematic diagram of a frequency domain noise reduction method using a short-time Fourier transform is shown in Figure 5.

Spectral subtraction
One of the classical methods for noise reduction is spectral subtraction [18, 92]. The method is based on a direct estimation of speech spectral
magnitude by subtracting the noise spectral magnitude from the noisy one. The attenuation filter for the spectral subtraction method can be formulated as

\[ H_{n[k]}^{\text{SS}} = \left( \max\left(\frac{|Y_{n[k]}|^r - |\hat{W}_{n[k]}|^r}{|Y_{n[k]}|^r}, 0 \right) \right)^{\frac{1}{r}}, \tag{26} \]

where \(|\hat{W}_{n[k]}|\) is an estimate of the noise spectrum amplitude, and \(r = 1\) for spectral magnitude subtraction [18] and \(r = 2\) for power spectral subtraction [92]. The power spectral subtraction method is closely related to the Wiener filtering approach, with the only difference in the power term (see section 2.2). For both magnitude and power subtractions, the phase of the noisy spectrum is used unprocessed in construction of the enhanced speech. One motivation is that the human auditory system is less sensitive to noise in spectral phase than in spectral magnitude [139], therefore the noise in spectral phase is less annoying.

Another formulation of the power spectral subtraction [92] is based on the complex Gaussian model (3) of the noisy spectrum. Under the Gaussian assumption, and due to the independence assumption, the noisy power spectrum of the \(k\)’th frequency bin is \(\lambda_{n[k]}^2[k] = \lambda_n^2[k] + \hat{\lambda}_n^2[k]\), where \(\lambda_n^2[k]\) and \(\hat{\lambda}_n^2[k]\) are the power spectra of speech and noise respectively. The PDF of \(Y_{n[k]}\) is given by

\[ f(Y_{n[k]}) = \frac{1}{\pi(\lambda_n^2[k] + \hat{\lambda}_n^2[k])} \exp\left(\frac{-|Y_{n[k]}|^2}{\lambda_n^2[k] + \hat{\lambda}_n^2[k]}\right). \tag{27} \]

Assuming that an estimate of the noise power spectrum \(\hat{\lambda}_n^2[k]\) exists, the maximum likelihood (ML) estimate of the speech power spectrum maximizes \(f(Y_{n[k]})\) with respect to \(\hat{\lambda}_n^2[k]\). Using the additional constraint that the power spectrum is nonnegative, the ML estimate of \(\lambda_n^2[k]\) is given by

\[ \hat{\lambda}_n^2[k] = \max(|Y_{n[k]}|^2 - \hat{\lambda}_n^2[k], 0). \tag{28} \]
The implementation of the spectral subtraction method is straightforward. However, the processed speech is plagued by the so-called musical noise phenomenon. The musical noise is the residual noise containing isolated tones located randomly in time and frequency. These tones correspond to spectral peaks in the original noisy signal and are left alone when their surrounding spectral bins in the time-frequency plane have been heavily attenuated. The musical noise is perceptually annoying and several methods have been derived from the spectral subtraction method targeting the musical noise problem [48, 83, 133, 136].

**Bayesian estimation approach**

Spectral subtraction was originally an empirically motivated method, and is often used in practice because of its simplicity. Alternatively, speech enhancement for noise reduction can be formulated in a statistical framework, such that prior knowledge of speech and noise is incorporated through prior PDFs. Using additionally a relevant cost function, speech can be estimated using the Bayesian estimation framework. The following sections, 2.2 and 2.3, are devoted to methods using the Bayesian estimation framework. In this subsection, a short introduction to Bayesian estimation is given.

At a conceptual level, assuming that speech and noise are random vectors distributed according to PDFs $f(x)$ and $f(w)$, the joint PDF of $x$ and $y$ can be formulated using the signal model (23). A Bayesian speech estimator minimizes the Bayes risk which is the expected cost with respect to both $x$ and $y$, defined as

$$\text{Bayes risk} = \int \int C(x, \hat{x}) f(x, y) dx dy,$$

(29)

where $C(x, \hat{x})$ is a cost function between $x$ and $\hat{x}$.

A commonly used cost function is the square error function,

$$C_{\text{MMSE}}(x, \hat{x}) = ||x - \hat{x}||^2,$$

(30)

and the optimal estimator minimizing the corresponding Bayes risk is the minimum mean square error (MMSE) estimator. The popularity of MMSE based estimation is due to its mathematical tractability and good performance. It can be shown that the MMSE speech estimator, e.g., [67], is the conditional expectation of speech given the noisy speech

$$\hat{x} = \int x f(x|y) dx.$$

(31)

**Noise estimation**

One key component of a single-channel noise reduction system is estimation of the noise statistics. Often, a speech estimator is designed assuming the
actual noise power spectrum is available, and its optimality is no longer ensured when the actual noise power spectrum is replaced by an estimate of the noise spectrum. Therefore, the performance of the speech estimator depends largely on the quality of the noise estimation algorithm.

A large number of noise estimation algorithms has been proposed. They are typically based on the assumption that 1) speech does typically not occupy the entire time-frequency plane 2) noise is “more stationary” than speech. By assuming that noise is stationary over a longer (than a frame) period of time, e.g., up to several seconds, the noise power spectrum can be estimated by averaging the noisy power spectra in the frequency bands with low or zero speech activity. The update of the noise estimate is commonly controlled through a voice-activity detector (VAD), e.g., [1,92], speech presence probability estimate [24, 87, 120], or order statistics [59, 88, 125]. For instance, the minimum statistics method [88] is based on the observation that the noisy power spectrum frequently approaches the noise power spectrum level due to speech absence in these frequency bins. To obtain the minimum statistics, a buffer is used to store past noisy power spectra of each frequency band. Bias compensation for taking into account the fluctuations of power spectrum estimation is proposed in [88]. In such a method, the size of the buffer is a trade-off. If the buffer is too small, it may not contain any speech inactive frames and speech may be falsely detected as noise. If the buffer is too large, it may not react fast enough for non-stationary noise types. In [88], the buffer length is set to 1.5 seconds.

2.2 Short-term model based approach

This section provides an overview of some Bayesian speech estimators using short-term models as discussed in section 1.1. In particular, we discuss Wiener filtering and spectral amplitude estimation methods.

Wiener filtering

Wiener filtering [141] is a classical signal processing technique that has been applied to speech enhancement. Wiener filters are the optimal linear time-invariant filters for the minimum mean square error (MMSE) criterion, and can be classified as Bayesian estimators under the constraint of linear estimation. The Wiener filter is also the optimal MMSE estimator under a Gaussian assumption.

Assuming linear estimation, the speech estimate in the time domain is of the form

$$\hat{x}_n = H_n y_n, \quad (32)$$
for a linear estimation matrix $H_n$. The MMSE linear estimator fulfills

$$H_n^{WF} = \arg \min_H \int \int ||x_n - Hy_n||^2 f(x_n, y_n) dx_n dy_n. \quad (33)$$

The optimal estimator can be solved by taking the first derivative of (33) with respect to $H$, and setting the obtained expression to zero. Using also the assumption that speech and noise have both zero mean and are statistically independent, the optimal linear estimator is

$$H_n^{WF} = \bar{D}_n (\bar{D}_n + \tilde{D}_n)^{-1}, \quad (34)$$

where $\bar{D}_n$ and $\tilde{D}_n$ are the covariance matrices of speech and noise, respectively, of frame $n$.

An alternative point of view for the Wiener filtering is through MMSE estimation under a Gaussian assumption. Assuming that speech and noise are of zero-mean Gaussian distributions, with covariance matrices $\bar{D}_n$ and $\tilde{D}_n$, it can be shown [67] that the conditional PDF of $x_n$ given $y_n$ is also Gaussian with mean $\bar{D}_n (\bar{D}_n + \tilde{D}_n)^{-1} y_n$. Hence, the MMSE estimator for a Gaussian model is linear and leads to the same result as the linear MMSE estimator of an arbitrary PDF (34).

Due to the correspondence between the convolution in the time domain and multiplication in the frequency domain, the Wiener filter in the frequency domain formulation has the simpler form of (25) with

$$H_n[k]^{WF} = \frac{\bar{\lambda}_n^2[k]}{\bar{\lambda}_n^2[k] + \tilde{\lambda}_n^2[k]}, \quad (35)$$

for the $k$'th frequency bin.

Practical implementation of the Wiener filter requires estimation of the second-order statistics (the power spectra) of speech and noise: $\bar{\lambda}_n^2[k]$ and $\tilde{\lambda}_n^2[k]$. The noise spectrum may be obtained through a noise estimation algorithm (section 2.1) and the speech spectrum may be estimated using (28). Hence, the actual implementation of the Wiener filtering is closely related to the power spectral subtraction method, and also suffers from the musical noise issue.

**Speech spectral amplitude estimation**

Another well-known and widely applied MMSE estimator for speech enhancement is the MMSE short-time spectral amplitude (STSA) estimator by Ephraim and Malah [38]. The STSA method does not suffer from musical noise in the enhanced speech, and is therefore interesting also from a perceptual point of view.

From the MMSE estimation result (31), the STSA estimator is formulated as the expected value of the speech spectral amplitude conditioned
on the noisy spectrum. The STSA estimator depends on two parameters, a-posteriori SNR $R_{\text{post}}^n[k]$ and a-priori SNR $R_{\text{prio}}^n[k]$, that need to be evaluated for each frequency bin. The a-posteriori SNR is defined as the ratio of the noisy periodogram and the noise power spectrum estimate$^2$

$$R_{\text{post}}^n[k] = \frac{|Y_n[k]|^2}{\hat{\lambda}_n^2[k]}, \quad \text{(36)}$$

and the a-priori SNR is defined as the ratio of the speech and noise power spectrum [38, 92]

$$R_{\text{prio}}^n[k] = \frac{\bar{\lambda}_n^2[k]}{\lambda_n^2[k]}, \quad \text{(37)}$$

The optimal attenuation filter for the STSA estimator is shown to be [38]

$$H_n[k] = \frac{\sqrt{\pi}}{2} \sqrt{\left( \frac{1}{R_{\text{post}}^n[k]} \right) \left( \frac{R_{\text{prio}}^n[k]}{1 + R_{\text{prio}}^n[k]} \right)} \cdot M \left( R_{\text{post}}^n[k] \right) \left( \frac{R_{\text{prio}}^n[k]}{1 + R_{\text{prio}}^n[k]} \right), \quad \text{(38)}$$

where $M$ is the function

$$M(\theta) = \exp \left( -\frac{\theta}{2} \right) (1 + \theta)I_0 \left( \frac{\theta}{2} \right) + \theta I_1 \left( \frac{\theta}{2} \right), \quad \text{(39)}$$

where $I_0$ and $I_1$ are the zeroth and first-order modified Bessel functions, respectively.

In practice, the a-priori SNR is not available and needs to be estimated. Ephraim and Malah proposed a solution based on the decision-directed approach, and the a-priori SNR is estimated as [38]

$$R_{\text{prio}}^n[k] = \alpha \frac{|H_{\text{prio}}^{-1}[k]Y_{\text{prio}}^{-1}[k]|^2}{\lambda_n^2[k]} + (1 - \alpha) \max(R_{\text{post}}^n[k] - 1, 0), \quad \text{(40)}$$

where $0.9 \leq \alpha < 1$ is a tuning parameter.

The STSA method has demonstrated superior perceptual quality compared to classic methods such as spectral subtraction and Wiener filtering. In particular, the enhanced speech does not suffer from the musical noise phenomenon. It has been argued [22] that the perceptual relevance of the method is mainly due to the decision-directed approach for estimating

$^2$This definition differs from the traditional definition of signal-to-noise ratio by including both speech and noise in the numerator.
The decision-directed approach has an attractive temporal smoothing behavior on the SNR estimate. It was shown [22] that for low $R_{\text{post}}$, i.e., the absence of speech, the $R_{\text{prio}}$ is an averaged SNR of a large number of successive frames; for high $R_{\text{post}}$, $R_{\text{prio}}$ follows $R_{\text{post}}$ with one-frame delay. This smoothing behavior has a significant effect on the perceptual quality of the enhanced speech, and the same strategy is applicable to other enhancement methods such as the Wiener filtering and spectral subtraction to reduce the musical noise in the enhanced speech.

Research challenges

The methods we have discussed in section 2.2 require an estimate of the noise power spectrum. Most noise estimation algorithms are based on the assumption that the noise is stationary or mildly non-stationary. Modern mobile devices are required to operate in acoustical environments with large diversity and variability in interfering noise. The non-stationarity may be due to the noise source itself, i.e., the noise contains impulsive or transient components, periodic patterns, or due to the movement of the noise source and/or the recording device. The latter scenario occurs commonly on a street, where noise sources (e.g., cars) move rapidly around.

Due to the increased degree of non-stationarity in such an environment, performance of most noise estimation algorithms is non-ideal and may significantly affect the performance of the speech estimation algorithm. Design of a noise reduction system that is capable of dealing with non-stationary noise is of great interest and a challenging task to the research community.

From the Bayesian estimation point of view, the performance of the speech estimator is determined by the accuracy of the speech and noise models, and perceptual relevancy of the distortion measure. Unfortunately, the true statistics of speech and noise are not explicitly available. Also, speech quality is a subjective measure that requires human listening experiments, and is expensive to assess. Objective measures that approximate the “true” perceptual quality have been proposed, e.g., for the evaluation of speech codecs [2, 14, 49, 50, 104, 110, 137, 138, 140], but most of them are complex and mathematically intractable for deriving a speech estimator. Nevertheless, a number of algorithms that are based on distortion measures that capture some properties of human perception have been proposed [57, 62, 79, 133, 136, 143]. For instance, the speech estimator in Paper A optimizes a criterion that allows for an adjustable level of residual noise in the enhanced speech, in order to avoid unnecessary speech artifacts.

Incorporating a higher degree of prior information of speech and noise can also improve the performance of the speech estimator. With short-term models, the assumed prior information is mainly the type of PDF, of which the parameters of both the speech and noise PDFs (their respective power spectra for Gaussian PDFs) need to be estimated. To incorporate stronger
prior information of speech and noise, more refined statistical models should be used. For instance, a more precise PDF of speech may be obtained by using a speech database and optimizing a long-term model through a training algorithm. In the next section, we discuss noise reduction methods using long-term models such as an HMM.

2.3 Long-term model based approach

This section provides an overview of noise reduction methods using long-term models from section 1.2. Long-term model based noise reduction systems include HMM based methods [35, 111], GMM based methods [32, 149] and codebook based methods [73, 121]. In this section, we focus mainly on methods based on hidden Markov models and their derivatives.

HMM-based speech estimation

Due to the inter-frame dependency of a HMM, the HMM based MMSE estimator (31) of speech can be formulated using current and past frames of the noisy speech,

$$\hat{x}_n = \int x_n f(x_n|Y_1^n) dx_n,$$

(41)

where \(Y_1^n = \{y_1, \ldots, y_n\}\) denotes the noisy observation sequence from frame 1 to \(n\). Due to the Markov chain assumption of the underlying state sequence, both the current and past noisy frames are utilized in the speech estimator of the current frame.

Assuming that both speech and noise are modeled using an HMM with Gaussian sub-models, it can be shown that the noisy model is a composite HMM with Gaussian sub-models due to the statistical independence of speech and noise. If the speech model has \(\bar{M}\) states and the noise model \(\tilde{M}\) states, the composite HMM has \(\bar{M}\tilde{M}\) states, one for each combination of speech and noise state. Under the Gaussian assumption, the noisy PDF of a given state is a Gaussian PDF with zero mean and covariance matrix \(D_s = \bar{D}_s + \tilde{D}_s\). The conditional PDF of speech given the noisy observation sequence can be formulated as

$$f(x_n|Y_1^n) = \sum_{s_n} f(s_n, x_n|Y_1^n)$$

$$= \sum_{s_n} f(s_n|Y_1^n) f(x_n|Y_1^n, s_n)$$

$$= \sum_{s_n} f(s_n|Y_1^n) f(x_n|y_n, s_n),$$

(42)

where the last step is due to the conditional independency property of an HMM, which states that the observation for a given state is independent of
other states and observations. Hence, the MMSE speech estimator using an HMM (41) can be formulated as

\[
\hat{x}_n = \sum_{s_n} f(s_n | Y^n) \int x_n f(x_n | y_n, s_n) dx_n \\
= \sum_{s_n} f(s_n | Y^n) \cdot \bar{D}_s (\bar{D}_s + \tilde{D}_s)^{-1} y_n,
\]

and each of the per-state conditional expectations is a Wiener filter (34).

The HMM based MMSE estimator is then a weighted sum of Wiener filters, where the weighting factors are the posterior state probabilities given the noisy observations. The weighting factors can be calculated efficiently using the forward algorithm [40, 105].

A significant difference between the HMM based estimator and the Wiener filter (34) is that the Wiener filter requires on-line estimation of both speech and noise statistics, while the per-state estimators of the HMM approach use speech model parameters obtained through off-line training (see section 1.2). Therefore, additional speech knowledge is incorporated in an HMM based method.

Noise reduction using an HMM was originally proposed by Ephraim in [35, 36] using an HMM with auto-regressive (AR) Gaussian sub-sources for speech modeling. The noise statistics are modeled using a single Gaussian PDF, and the noise model parameters are obtained using the first few frames of the noisy speech that are considered to be noise-only. Alternatively, an additional noise estimation algorithm may be used to obtain the noise statistics. In [111], the HMM based methods are extended to allow HMM based noise models. The method of [111] uses several noise models, obtained off-line for different noise environments, and a classifier is used to determine which noise model must be used. By utilizing off-line trained models for both speech and noise, strong prior information of speech and noise is incorporated. By having more than one noise state, each representing a distinct noise spectrum, the method can handle rapid changes of the noise spectrum within a specific noise environment, and can provide significant improvement for non-stationary noise environments.

**Stochastic gain modeling**

A key issue of using long-term speech/noise models obtained through off-line training is the energy mismatch between the model and the observations. Since the model is obtained off-line using training data, the energy level of the training data may differ from the speech in the noisy observation, causing a mismatch. Hence, strategies for gain adaptation are needed when long-term models are used.
When using an AR-HMM for speech modeling [35], the signal is modeled as an AR process in each state. The excitation variance of the AR model is a parameter that is obtained from the training data. Consequently, the different energy levels of a speech sound, typically due to differences in pronunciation and/or different vocalizations of individual speakers, are not efficiently modeled. In fact, an AR-HMM implicitly assumes that speech consists of a finite number of frame energy levels. Another energy mismatch between the speech model and the observation is due to different recording conditions during off-line training and on-line estimation, and may be considered as a global energy mismatch. In the MMSE speech estimator of [35], a global gain factor is used to reduce this mismatch.

A similar problem appears in noise modeling, since the noise energy level in the noisy environment is inherently unknown, time-varying, and in most natural cases, different from the noise energy level in the training. In [111], a heuristic noise gain adaptation using a voice activity detector has been proposed, where the adaptation is performed in speech pauses longer than 100 ms. In [146], continuous noise gain model estimation techniques were proposed.

In Paper A, we propose a new HMM based gain-modeling technique that extends the HMM-based methods [35, 111] by introducing explicit modeling of both speech and noise gains in a unified framework. The probability density function of \( x_n \) for a given state \( \bar{s} \) is the integral over all possible speech gains. Modeling the speech energy in the logarithmic domain, we then have

\[
f_{\bar{s}}(x_n) = \int_{-\infty}^{\infty} f_{\bar{s}}(g'_n) f_{\bar{s}}(x_n|g'_n) dg'_n, \quad (44)
\]

where \( g'_n = \log \bar{g}_n \) and \( \bar{g}_n \) denotes the speech gain in the linear domain.

The extension of Paper A over the traditional AR-HMM is the stochastic modeling of the speech gain \( \bar{g}_n \), where \( \bar{g}_n \) is considered as a stochastic process. The PDF of \( \bar{g}_n \) is modeled using a state-dependent log-normal distribution, motivated by the simplicity of the Gaussian PDF and the appropriateness of the logarithmic scale for sound pressure level. In the logarithmic domain, we have

\[
f_{\bar{s}}(g'_n) = \frac{1}{\sqrt{2\pi\sigma^2_{\bar{s}}}} \exp\left( -\frac{1}{2\sigma^2_{\bar{s}}} (g'_n - \bar{\mu}_{\bar{s}} - \bar{\eta}_n)^2 \right), \quad (45)
\]

with mean \( \bar{\mu}_{\bar{s}} + \bar{\eta}_n \) and variance \( \sigma^2_{\bar{s}} \). The time-varying parameter \( \bar{\eta}_n \) denotes the speech-gain bias, which is a global parameter compensating for the long-term energy level of speech, e.g., due to a change of physical location of the recording device. The parameters \( \{\bar{\mu}_{\bar{s}}, \sigma^2_{\bar{s}}\} \) are modeled to be time-invariant, and can be obtained off-line using training data, together with the other speech HMM parameters.
The model proposed in Paper A is referred to as a stochastic-gain HMM (SG-HMM), and may be applied to both speech and noise. To obtain the time-invariant parameters of the model, an off-line training algorithm is proposed based on the EM technique. For time-varying parameters, such as $q_n$, an on-line estimation algorithm is proposed based on the recursive EM technique. In Paper A, we demonstrate through objective and subjective evaluations that the new model improves the modeling of speech and noise, and also the noise reduction performance.

The proposed HMM based gain modeling is related to the codebook based methods [73, 74, 122, 124]. In the codebook methods, the prior speech and noise models are represented as codebooks of spectral shapes, represented by linear prediction (LP) coefficients. For each frame, and each combination of speech and noise code vectors, the excitation variances are estimated instantaneously. In [73, 123], the most likely combination of code vectors are used to form the Wiener filter. In [74, 124], a weighted sum of Wiener filters is proposed, where the weighting is determined by the posterior probabilities similarly to (43). The Bayesian formulation in [124] is based on the assumption that both the code vectors and the excitation variances are uniformly distributed.

**Adaptation of the noise HMM**

It is widely accepted that the characteristics of speech can be learned from a (sufficiently rich) database of speech. A successful example is in the speech coding application, where, e.g., the linear prediction coefficients (LPC) can be represented with transparent quality using a finite number of bits [97]. However, noise may vary to a large extent in real-world situations. It is unrealistic to assume that one can capture all of this variation in an initial learning stage, and this implies that on-line adaptation to changing noise characteristics is necessary.

Several methods have been proposed to allow noise model adaptation based on different assumptions, e.g., [149] for a GMM and [84, 93] for AR-HMM. The algorithm in [84, 149] only applies to a batch of noisy data, e.g., one utterance, and is not directly applicable for on-line estimation. The noise model in [149] is limited to stationary Gaussian noise (white or colored). In [84, 93], a noise HMM is estimated from the observed noisy speech, using a voice-activity-detector (VAD), such that noise-only frames are used in the on-line learning.

In Paper B, we propose an adaptive noise model based on SG-HMM. The results from Paper A demonstrate good performance for modeling speech and noise. The work is extended in Paper B with an on-line noise estimation algorithm for enhancement in unknown noise environments. The model parameters are estimated on-line using the noisy observations without requiring a VAD. Therefore, the on-line learning of the noise model is continuous.
and facilitates adaptation and correction to changing noise characteristics. Estimation of the noise model parameters is based on maximizing the likelihood of the noisy model, and the proposed implementation is based on the recursive expectation maximization (EM) framework [72,132].

3 Flexible source coding

The second part of this thesis focuses on flexible source coding for unreliable networks. In Paper C, we propose a quantizer design based on a Gaussian mixture model for describing the source statistics. The proposed quantizer design allows for rate adaptation in real-time depending on the current network condition. This section provides an overview of the theory and practice that form the basis for quantizer design.

The section is organized as follows. Basic definitions and design criteria are formulated in section 3.1. High-rate theory is introduced in section 3.2. In section 3.3, some practical coder designs motivated by high-rate theory are discussed.

3.1 Source coding

Information theory, founded by Shannon in his legendary 1948 paper [116], forms the theoretical foundation for modern digital communication systems. As shown in [116], the communication chain can roughly be divided into two separate parts: source coding for data compression and channel coding for data protection. Optimal performance of source and channel coding can be predicted using the source and channel coding theorems [116]. An implication of the theorems is that optimal transmission of a source may be achieved by separately performing source coding followed by channel coding, a result commonly referred to as the source-channel separation theorem.

The source and channel coding theorems motivate the design of essentially all modern communication systems. The optimality of separate source and channel coding is based on the assumption of coding in infinite dimensions, therefore requires an infinite delay. While the optimality is not guaranteed in any practical coders with finite delay, the design is used in practical applications due to its simplicity.

Source coding consists of lossless coding [63,75–77,109,142], to achieve compression without introducing error, and lossy coding [16,117], which achieves a higher compression level but introduces distortion. Modern source coders utilize lossy coding and remove both irrelevancy and redundancy. They often incorporate lossless coding methods to remove redundancy from parameters that require perfect reconstruction, such as quantization indices. Such a system typically consists of a quantizer (lossy coding) followed by an entropy coder (lossless coding). The rate-distortion perfor-
formance for the optimal coder design can be predicted by rate-distortion theory [116]. The theory provides a theoretical bound for the achievable rate without exceeding a given distortion. The rate-distortion bound demonstrates an interesting and intuitive behavior, namely that the achievable rate is a monotonically decreasing convex function of distortion. An example of a rate-distortion function for a unit-variance Gaussian source is shown in Figure 6. While the theory is derived under the assumption of infinite dimensions, it is expected that the behavior is also relevant in lower dimensions. Flexible source coder design should allow adjustment of the rate-distortion trade-off, e.g., multiple operation points on the rate-distortion curve, such that the coder can operate optimally over a large range of rates. For an unreliable network, it is then possible to adapt the rate of the coder in real-time to the current network condition. By operating at a rate adapted to the network condition, the coder performs with the best quality (lowest distortion) for the available bandwidth.

Quantization

Let \( \mathbf{x} \) denote a source vector in the \( K \) dimensional Euclidean space \( \mathbb{R}^K \) distributed according to the PDF \( f(\mathbf{x}) \). Quantization is a function

\[
\hat{x} = Q(\mathbf{x})
\]

(46)

that maps each source vector \( \mathbf{x} \in \mathbb{R}^K \) to a code vector \( \hat{x} \) from a codebook, which consists of a countable set of vectors in \( \mathbb{R}^K \). Each code vector is associated with a quantization cell (decision region), \( \text{cell}(\hat{x}) \), defined as

\[
\text{cell}(\hat{x}) = \{ \mathbf{x} \in \mathbb{R}^K : Q(\mathbf{x}) = \hat{x} \}.
\]

(47)
An index is assigned to each code vector in the codebook, and the index is coded with a codeword, i.e., a sequence of bits. Each quantization index corresponds to a particular codeword and all codewords together form a uniquely decodable code, which implies that any encoded sequence of indices can be uniquely decoded. For $K = 1$, the quantizer applies to scalar values, and is called a scalar quantizer. The more general quantizer with $K \geq 1$ is called a vector quantizer.

**Rate**

Rate can be defined as the average codeword length per source vector. Since the available bandwidth is limited, the quantizer is optimized under a rate constraint. The codewords may be constrained to have a particular fixed length, resulting in a fixed-rate coder, or constrained to have a particular average length, resulting in a variable-rate coder.

For a fixed-rate coder with $N$ code vectors in the codebook, the rate is given by

$$R_{\text{fix.}} = \lceil \log_2 N \rceil,$$

(48)

where $\lceil \cdot \rceil$ denotes the ceiling function (the smallest integer equal to or larger than the input value). To achieve efficient quantization, $N$ is typically chosen to be an integer power of 2, to avoid rounding.

For a variable-rate coder, code vectors are coded using codewords with different lengths. The rate determines the average codeword length,

$$R_{\text{var.}} = \sum_{\hat{x}} p(\hat{x}) \ell(\hat{x}),$$

(49)

where $\ell(\hat{x})$ denotes the codeword length of $\hat{x}$ and $p(\hat{x}) = \int_{\text{cell}(\hat{x})} f(x) dx$ is the probability mass function of $\hat{x}$, obtained by integrating $f(x)$ over the quantization cell of $\hat{x}$. From the source coding theorem [116], the optimal variable-rate code has an average rate approaching the entropy of $\hat{x}$, denoted $H$,

$$H = -\sum_{\hat{x}} p(\hat{x}) \log_2 p(\hat{x}).$$

(50)

Using an entropy code such as the arithmetic code [75–77, 109, 142], this rate can be approached arbitrarily closely by increasing the sequence length.

**Distortion**

The performance of a quantizer is measured by the difference or error between the source vector and the quantized code vector. This difference defines an error function, denoted $d(x, \hat{x})$, that quantifies the distortion due to the quantization of $x$. The performance of the system is given by the
average distortion defined as the expected error with respect to the input variable,

\[ D = \int d(x, \hat{x}) f(x) dx. \quad (51) \]

The error function should be meaningful with respect to the source, and it typically fulfills some basic properties, e.g., \[53\]

- \( d(x, \hat{x}) \geq 0 \)
- if \( x = \hat{x} \) then \( d(x, \hat{x}) = 0 \).

The most commonly used error function is the square error function,

\[ d(x, \hat{x}) = \frac{1}{K} ||x - \hat{x}||^2 = \frac{1}{K} (x - \hat{x})^T (x - \hat{x}), \quad (52) \]

where \( || \cdot || \) denotes the \( L_2 \) vector norm, and \((\cdot)^T\) denotes the transpose. The mean distortion of (52) is referred to as the mean square error (MSE) distortion. The MSE distortion is commonly used due to its mathematical tractability.

Other distortion measures may be conceptually or computationally more complicated, e.g., by incorporating perceptual properties of the human auditory systems. If a complex error function can be assumed to be continuous and have continuous derivatives, the error function can be approximated under a small error assumption. By applying a Taylor series expansion of \( d(x, \hat{x}) \) around \( x = \hat{x} \), and noting that the first two expansion terms vanish for \( x = \hat{x} \), the error function may then be approximated as \[44\]

\[ d(x, \hat{x}) \approx \frac{1}{2K} (x - \hat{x})^T W(\hat{x})(x - \hat{x}), \quad (53) \]

where \( W(\hat{x}) \) is a \( K \times K \) dimensional matrix, the so-called sensitivity matrix, with the \( i, j \)th element defined by

\[ W(\hat{x})[i, j] = \left. \frac{\partial^2 d(x, \hat{x})}{\partial x[i] \partial x[j]} \right|_{x=\hat{x}}, \quad (54) \]

is the Jacobian matrix evaluated at the expansion point \( \hat{x} \). In [44], the sensitivity matrix approach was applied to quantization of speech linear prediction coefficients (LPC) using, e.g., the log-spectral distortion (LSD) measure. In [101, 102], a sensitivity matrix was derived and analyzed for a perceptual distortion measure based on a sophisticated spectral-temporal auditory model [29, 30].
Quantizer design

The optimal quantizer can be formulated using the constrained optimization framework, by minimizing the expected distortion, \( D \), while constraining the average rate below or equal to the available rate budget. The constraint is defined using (48) for a fixed-rate coder or (50) for a variable-rate coder. The resulting quantizer designs are referred to as a resolution-constrained quantizer and an entropy-constrained quantizer, respectively. In general, an entropy-constrained quantizer has better rate-distortion performance compared to a resolution-constrained quantizer, due to the flexibility of assigning bit sequences of different lengths to different code vectors according to the probability of their appearance. In Paper C, we propose a flexible design of an entropy-constrained vector quantizer.

Traditional quantizer design is typically based on an iterative training algorithm [23, 81, 82, 91]. By iteratively optimizing the encoder (decision regions or quantization cells) and decoder (reconstruction point of each cell), and ensuring that the performance is non-decreasing at each iteration, a locally optimal quantizer is constructed. Using a proper initialization scheme, and given sufficient training data, a good rate-distortion performance can be achieved by a training based quantizer.

An issue with the traditional quantizer design is the computational complexity and memory requirement associated with the codebook search and storage. As the number of code vectors is exponentially increasing with respect to the rate, and in general a higher rate (per vector) is needed for quantizing higher dimensional vectors, the approach is often limited to low-rate and scalar or low-dimensional vector quantizers.

Another disadvantage of the traditional training based quantizer design is that the resulting quantizer has a fixed codebook. Its rate-distortion performance is determined off-line, and adaptation of the codebook according to the current available bandwidth is unfeasible without losing optimality. To allow a certain flexibility, one can optimize multiple codebooks, each for a different rate, and select the best codebook. A disadvantage of this approach is, however, the additional memory requirement, e.g., for storage of multiple codebooks.

It is desirable to have a flexible quantizer design that allows for a fine-grained rate adaptation, such that the rate can be selected from a larger range with a finer resolution, without increasing the memory requirement. To achieve that, the quantizer design should avoid usage of an explicit codebook (that requires storage), and the code vectors are constructed on-line for a given rate. High-rate theory provides a theoretical basis for designing such a quantizer.
3.2 High-rate theory

High-rate theory is a mathematical framework for analyzing the behavior and predicting the performance of a quantizer of any dimensionality. High-rate theory originated from the early works on scalar quantization [15, 47, 96, 98], and was later extended for vector quantization [45, 144]. A comprehensive tutorial on quantization and high-rate theory is given in [54]. Textbooks treating the topic include [16, 46, 56, 71].

High-rate theory applies when the rate is high with respect to the smoothness of the source PDF $f(x)$. Under this high-rate assumption, $f(x)$ may be assumed to be constant within each quantization cell, such as $f(x) \approx f(\hat{x})$ in $\text{cell}(\hat{x})$. The probability mass function (PMF) of $\hat{x}$ can then be approximated using

$$p(\hat{x}) \approx f(\hat{x}) \text{vol}(\hat{x}),$$

where $\text{vol}(\hat{x})$ denotes the volume of $\text{cell}(\hat{x})$.

Instead of using a codebook, a high-rate quantizer can be formulated as a quantization point density function, specifying the distribution of the code vectors, e.g., the number of code vectors per unit volume. A quantization point density function may be defined as a continuous function that approximates the inverse volumes of the quantization cells [71],

$$g_c(x) \approx \frac{1}{\text{vol}(x)} \text{, if } x \in \text{cell}(\hat{x}).$$

Under the high-rate assumption, optimal solutions of $g_c(x)$ can be solved for different rate constraints. Practical quantizers designed according to the obtained $g_c(x)$ approach optimality with increasing rate. Nevertheless, quantizers designed based on the high-rate theory have shown competitive performance also at a lower rate. In Paper C, the actual performance of the proposed quantizer is very close to the predicted performance by the high-rate theory for rates higher than 3-4 bits per dimension.

High-rate distortion

The average distortion of a quantizer can be reformulated using the high-rate assumption. For the MSE distortion measure, (51) can be written as

$$D_{\text{MSE}} = \frac{1}{K} \int \|x - \hat{x}\|^2 f(x) dx$$

$$\approx \frac{1}{K} \sum_{\hat{x}} \int_{\text{cell}(\hat{x})} \|x - \hat{x}\|^2 f(\hat{x}) dx$$

$$= \frac{1}{K} \sum_{\hat{x}} \frac{p(\hat{x})}{\text{vol}(\hat{x})} \int_{\text{cell}(\hat{x})} \|x - \hat{x}\|^2 dx.$$
Let
\[ C(\hat{x}) = \frac{1}{K} \text{vol}(\hat{x})^{-\frac{K+2}{K}} \int_{\text{cell}(\hat{x})} ||x - \hat{x}||^2 dx, \] 
(58)
denote the coefficient of quantization of \( \hat{x} \). \( C(\hat{x}) \) is the normalized distortion with respect to volume, and shows the effect of the geometry of the cell. The distortion can then be written as
\[ D_{\text{MSE}} = \sum_{\hat{x}} p(\hat{x}) \text{vol}(\hat{x})^{\frac{3}{2}} C(\hat{x}). \] 
(59)

Gersho conjectured that the optimal quantizer has the same cell geometry for all quantization cells, and has a constant coefficient of quantization \( C_{\text{opt}} \) \cite{45}. Using the optimal cell geometry, \( C(\hat{x}) \approx C_{\text{opt}} \), then the distortion is
\[ D_{\text{MSE}} \approx C_{\text{opt}} \sum_{\hat{x}} p(\hat{x}) \text{vol}(\hat{x})^{\frac{3}{2}} \] 
\[ \approx C_{\text{opt}} \int f(x) g_c(x)^{-\frac{2}{K}} dx. \] 
(60)

Resolution-constrained vector quantizer

The optimal resolution-constrained vector quantizer (RCVQ) has the quantization point density function \( g_c(\cdot) \) that minimizes the average distortion \( D \), under a rate constraint that the number of code vectors equals \( N \). The constraint may be formulated through integration of \( g_c(x) \),
\[ \int g_c(x) dx = N. \] 
(61)

The optimal \( g_c(\cdot) \) minimizes the extended criterion
\[ \eta_{\text{rcvq}} = C_{\text{opt}} \int f(x) g_c(x)^{-\frac{2}{K}} dx + \lambda \left( \int g_c(x) dx - N \right), \] 
(62)
where the first term is the high-rate distortion (60) and \( \lambda \) is the Lagrange multiplier for the rate constraint. Solving the Euler-Lagrange equation, and combining with (61), the optimal \( g_c(x) \) for constrained-resolution vector quantization with respect to the MSE distortion measure is given by
\[ g_c(x) = \frac{N f(x)^{\frac{K}{K+2}}}{\int f(x)^{\frac{K}{K+2}} dx}. \] 
(63)
Entropy-constrained vector quantizer

The optimal entropy-constrained vector quantizer (ECVQ) has a quantization point density function $g_c(\cdot)$ that minimizes the average distortion $D$, under the constraint that the entropy of the code vectors equals a desired target rate $R_t$. The entropy of the quantized variable at a high rate can be formulated as

$$H \approx -\sum_x f(\hat{x}) \text{vol}(\hat{x}) \log_2 f(\hat{x}) \text{vol}(\hat{x})$$

$$\approx -\int f(x) \log_2 \frac{f(x)}{g_c(x)} dx$$

$$= h + \int f(x) \log_2 g_c(x) dx,$$

(64)

where $h = -\int f(x) \log_2 f(x) dx$ is the differential entropy of the source. Therefore, the constraint can be written as $\int f(x) \log_2 g_c(x) dx = R'_t$, where $R'_t = R_t - h$. The optimal $g_c(\cdot)$ minimizes the extended criterion

$$\eta_{ecvq} = C_{opt} \int f(x) g_c(x) - \frac{1}{2} dx + \lambda \left( \int f(x) \log_2 g_c(x) dx - R'_t \right),$$

(65)

where $\lambda$ is the Lagrange multiplier for the entropy constraint.

The optimal $g_c(x)$ can be solved similarly to the RCVQ case, and the resulting $g_c(x)$ is a constant $[45]$,

$$g_c(x) = g_c = 2^{R_t - h}.$$

(66)

This important result implies that, at high rate, uniformly distributed quantization points form an optimal quantizer if the point indices are subsequently coded using an entropy code.

3.3 High-rate quantizer design


Lattice quantization

A lattice $\Lambda$ is an infinite set of regularly arranged vectors in the Euclidean space $\mathbb{R}^K$ $[28]$. It can be defined through a generating matrix $G$, assumed to be a square matrix,

$$\Lambda(G) = \{G^T u : u \in \mathbb{Z}^k\},$$

(67)
where $G$ is a matrix with linearly independent row vectors that form a set of basis vectors in $\mathbb{R}^K$. The lattice points generated through $G$ are integer linear combinations of the row vectors of $G$.

Applied to quantization, each lattice point $\lambda \in \Lambda$ is associated with a Voronoi region containing all points in $\mathbb{R}^K$ closest to $\lambda$ according to the Euclidean distance measure. The Voronoi region corresponds to the decision region of a quantizer with the MSE distortion measure. Hence, a closest point search algorithm can be used for quantization. Figure 7 shows a two-dimensional lattice, the hexagonal lattice (defined by $G = \begin{bmatrix} 1 & 0 \\ 0.5 & \sqrt{3}/2 \end{bmatrix}$), together with the Voronoi cells.

The high-rate theory for entropy-constrained vector quantization forms the theoretical basis for using uniformly distributed code vectors. Such vectors may be generated using a lattice, and the quantizer is then a lattice quantizer. The design of a coder based on an entropy-constrained lattice quantizer consists of the following steps: 1) selection of a suitable lattice for quantization, 2) design of a closest point search algorithm for the selected lattice, and 3) indexing and assigning variable-length codewords to the code vectors.

The selection of a good lattice for quantization is considered in [9] for the MSE distortion. Since the MSE performance of a lattice quantizer is proportional to the coefficient of quantization associated with the Voronoi cell shape of the lattice, lattice optimization for quantization is based on an optimization with respect to the coefficient of quantization. A study and tutorial on the “best known” lattices for quantization is given in [9]. A numerical optimization method for obtaining a locally optimal lattice for quantization in an arbitrary dimension is proposed in [9].

Closest point search algorithms for lattices are considered in [10, 27]. The algorithms proposed in [27] are based on the geometrical structures...
of the lattices, and are therefore lattice-specific. A general searching algorithm that applies to an arbitrary lattice has been proposed [10]. Both methods are considerably faster than a full search over all code vectors in a traditional codebook, by utilizing the geometrically regular arrangement of code vectors. As the search complexity and memory requirement using a traditional codebook are exponentially increasing with respect to the vector dimensionality and rate, lattice quantization is an attractive approach for high-dimensional vector quantization.

The practical application of entropy-constrained lattice quantization has been limited due to lack of a practical algorithm for indexing and assigning variable-length codewords that work at high rates and dimensions. In [25, 26, 66], lattice truncation and indexing the lattice vectors are considered. However, designing a practical entropy code is still limited to low dimensions and rates, due to the exponentially increasing number of code vectors in the truncated lattice. In Paper C, we propose a practical indexing and entropy coding algorithm for a lattice quantization of a Gaussian variable as part of a GMM based entropy-constrained vector quantizer design.

**GMM-based quantization**

High-rate theory assumes that the PDF of the source is known. However, for most practical sources, the PDF is unknown and needs to be estimated from the data. As discussed in section 2.3, GMMs form a convenient tool for modeling real-world source vectors with an unknown PDF [107]. They have been successfully applied to quantization, e.g., [52, 58, 114, 127, 128, 131, Paper C].

GMMs have been applied to approximate the high-rate optimal quantization point density for resolution-constrained quantization [58, 114]. A codebook can be generated using a pseudo random vector generator, once the optimal quantization point density is obtained. Thus, this method does not require the storage of multiple codebooks for different rate requirements. However, the method shares the same search complexity as a traditional codebook based quantizer. The application is therefore limited to low rates and dimensions. Nevertheless, the method is useful for predicting the performance of an optimal quantizer [58, 114].

Another GMM quantizer design is based on classification and component quantizers [127–129]. Such a quantizer is based on quantization of the source vector using only one of the mixture component quantizers. The index of the quantizing mixture component is transmitted as side information. The scheme is conceptually simple, since the problem is reduced to quantizer design for a Gaussian component. Similar ideas based on classified quantization appears in image coding [131] and speech coding [68–70].

The methods of [127–129] are resolution-constrained vector quantizers optimized for a fixed target rate. The overall bit rate budget is distributed
among the mixture components and for transmitting the component index. The component quantizers are based on the so-called companding technique. The companding technique attempts to approximate the optimal high-rate codebook of a Gaussian variable. It consists of an invertible mapping function that maps the input variable into a domain where a lattice quantizer is applied. The forward transform is referred to as the compressor and the inverse transform as expander. Applying the expander function to the lattice structured codebook would result in a codebook that approximates the optimal codebook in the original domain. Unfortunately, the companding technique is suboptimal for vector dimensions higher than two \([20, 21, 45, 95, 119]\). The sub-optimality is due to the fact that the expander function scales the quantization cells differently in different dimensions and in different locations in space. Hence, the quantization cells lose their optimal cell shape in the original domain. The great advantage of the schemes proposed in \([127–129]\) is the low computational complexity. Also, codebook storage is not required by utilizing a lattice quantizer in the transformed domain.

Applying classified quantization for entropy constrained quantization has been discussed in \([52, 71]\). Motivated by high-rate theory, the component quantizers of an entropy constrained quantizer are based on lattice structured codebooks. The focus of \([52, 71]\) is on the high-rate analysis of such a quantizer. In Paper C, we propose a practical GMM-based entropy constrained quantizer using the classified quantization framework. In particular, a practical solution to indexing and entropy coding for the lattice quantizers of the Gaussian components is proposed. We show that, under certain conditions, the scheme approaches the theoretically optimal performance with increasing rate.

4 Summary of contributions

The focus of this thesis is on statistical model based techniques for speech enhancement and coding. The main contributions of the thesis can be summarized as

- improved speech and noise modeling and estimation techniques applied to speech enhancement, and

- improvements towards flexible source coding.

The thesis consists of three research articles. Short summaries of the articles are presented below.

4 Summary of contributions

The focus of this thesis is on statistical model based techniques for speech enhancement and coding. The main contributions of the thesis can be summarized as

- improved speech and noise modeling and estimation techniques applied to speech enhancement, and

- improvements towards flexible source coding.

The thesis consists of three research articles. Short summaries of the articles are presented below.
Paper A: HMM-based Gain-Modeling for Enhancement of Speech in Noise

In Paper A, we propose a hidden Markov model (HMM) based speech enhancement method using stochastic-gain HMMs (SG-HMMs) for both speech and noise. We demonstrate that accurate modeling and estimation of speech and noise gains facilitate good performance in speech enhancement. Through the introduction of stochastic gain variables, energy variation in both speech and noise is explicitly modeled in a unified framework. The speech gain models the energy variations of the speech phones, typically due to differences in pronunciation and/or different vocalizations of individual speakers. The noise gain helps to improve the tracking of the time-varying energy of non-stationary noise. The expectation-maximization (EM) algorithm is used to perform off-line estimation of the time-invariant model parameters. The time-varying model parameters are estimated on-line using the recursive EM algorithm. The proposed gain modeling techniques are applied to a novel Bayesian speech estimator, and the performance of the proposed enhancement method is evaluated through objective and subjective tests. The experimental results confirm the advantage of explicit gain modeling, particularly for non-stationary noise sources.


In Paper B, we extend the work in Paper A and introduce an on-line noise estimation algorithm. A stochastic-gain HMM is used to model the statistics of non-stationary noise with time-varying energy. The noise model is adaptive and the model parameters are estimated on-line from noisy observations. The parameter estimation is derived for the maximum likelihood criterion and the algorithm is based on the recursive expectation maximization (EM) framework. The proposed method facilitates continuous adaptation to changes of both noise spectral shapes and noise energy levels, e.g., due to movement of the noise source. Using the estimated noise model, we also develop an estimator of the noise power spectral density (PSD) based on recursive averaging of estimated noise sample spectra. We demonstrate that the proposed scheme achieves more accurate estimates of the noise model and noise PSD. When integrated to a speech enhancement system, the proposed scheme facilitates a lower level of residual noise in the enhanced speech.

Paper C: Entropy-Constrained Vector Quantization Using Gaussian Mixture Models

In Paper C, a flexible entropy-constrained vector quantization scheme based on Gaussian mixture models (GMMs), lattice quantization, and arithmetic
coding is presented. A source vector is assumed to have a probability density function described by a GMM. The source vector is first classified to one of the mixture components, and the Karhunen-Loève transform of the selected mixture component is applied to the vector, followed by quantization using a lattice structured codebook. Finally, the scalar elements of the quantized vector are entropy coded sequentially using a specially designed arithmetic code. The proposed scheme has a computational complexity that is independent of rate, and quadratic with respect to vector dimension. Hence, the scheme facilitates quantization of high dimensional source vectors. The flexible design allows for changing of the average rate on-the-fly. The theoretical performance of the scheme is analyzed under a high-rate assumption. We show that, at high rate, the scheme approaches the theoretically optimal performance, if the mixture components are located far apart. The performance loss when the mixture components are located closely to each other can be quantified for a given GMM. The practical performance of the scheme was evaluated through simulations on both synthetic and speech-derived source vectors, and competitive performance has been demonstrated.

References


