Graz University of Technology

M.Sc. Thesis

Recent Advances in Knowledge-based Speech Enhancement

conducted at the Signal Processing and Speech Communications Laboratory
Graz University of Technology, Austria

by Christian Stetco, B.Sc.

Supervisors: Pejman Mowlaee, PhD

Examiners:

Graz, May 14, 2018

Acknowledgement

To my beloved family.

Abstract

Speech enhancement systems which rely on conventional methods show outstanding performance in stationary noise scenarios but suffer in non-stationary environments which are present in everyday life situations. Recent studies were able to circumvent such issues by storing a priori information of speech. Therefore, spectral shapes computed from linear predictive coefficients were obtained as codebooks and full-search algorithms were implemented finding the best match among all spectral shapes. This thesis presents a new approach by replacing the codebooks with Gaussian Mixture Models leading to a uncertainty model and soft clusters among spectral shapes. Furthermore, the use of variational Bayesian methods allows a full probabilistic description by computing approximations of the a posteriori distributions of interest. The approximation is obtained in an iterative fashion using an EM-based algorithm for improving speech on a frame-by-frame basis. Experiments were accomplished to give a deeper understanding of the new approach and the performance was evaluated by means of speech quality and intelligibility measures.

Kurzfassung

Konventionelle Sprachsignalverbesserungssysteme zeigen herausragende Ergebnisse in stationären Rauschszenarien scheitern jedoch in nicht stationären Umgebungen, die sich im alltäglichen Leben wieder finden. Jüngste Studien konnten diese Probleme beheben, indem a priori Information von Sprachsignalen gespeichert wurde. Um dies zu erzielen, wurden die Einhüllenden der spektralen Leistungsdichte von autoregressiven Koeffizienten berechnet und als Codebücher gespeichert und Full-Search Algorithmen wurden implementiert um die besten Treffer zu finden. In dieser Arbeit wird ein neuer Ansatz präsentiert indem die Codebücher durch Gaus'sche Mischverteilungen ersetzt werden welche eine Unsicherheitsmodellierung und weiche Gruppierung ermöglicht. Außerdem erlauben variationelle Bayes'sche methoden eine volle probabilistische Beschreibung indem Approximationen der a posteriori Verteilungen berechnet werden. Die Approximationen werden in einer iterativen Weise auf Basis eines EM Algorithmus berechnet, welches zu einer segmentbasierenden Verbesserung von Sprachsignalen führt. Experimente wurden durchgeführt um einen tiefen Einblick in die Methode zu gewährleisten und die Evaluierung wurde mittles Maße für Sprachqualität und Sprachverständlichkeit ermittelt.

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

....

date (signature)

List of Abbreviations

Abbreviation	Description
AR	Autoregressive
BN	Bayesian Network
CB	Codebook
EM	Expectation Maximization
GLA	generalized Llody algorithm
GMM	Gaussian Mixture Model
i.i.d.	independent and identical distributed
IS	Itakuro-Saito
KF	Kalman Filter
LDS	Linear Dynamical System
LSA	Log
MAP	Maximum a posteriori
ML	Maximum Likelihood
MMSE	Minimum Mean Square Error
pdf	probability density function
PESQ	Perceptual Evaluation of Speech Quality
SNR	Signal-to-Noise ratio
SSNR	segmental Signal-to-Noise ratio
STOI	Short Time Objective Intelligibility
STP	short-term predictor
VB	Variational Bayesian
VB-EM	Variational Bayesian Expectation Maximization
VQ	Vector Quantization

Mathematical Notation

Notation	Description
$x[n]$	clean speech signal
$y[n]$	noisy speech signal
$w[n]$	noise signal
$\boldsymbol{a}_x$	speech LP coefficients
$\boldsymbol{a}_w$	noise LP coefficients
N_x	number of speech mixtures
N_w	number of noise mixtures
$\boldsymbol{z}$	hidden state variables
$\boldsymbol{\theta}_x$	speech STPs
$\boldsymbol{\theta}_w$	noise STPs
σ_x^2	speech excitation variance
σ_w^2	noise excitation variance
$\boldsymbol{R}_x$	speech covariance matrix
$\boldsymbol{R}_w$	noise covariance matrix
$S_{xx}(\omega)$	speech power spectrum
$S_{ww}(\omega)$	noise power spectrum
$\overline{S}_{xx}(\omega)$	normalized speech power spectrum
$\overline{S}_{ww}(\omega)$	normalized noise power spectrum
ω	normalized frequency variable
N	framelength
P	speech LP order
Q	noise LP order
$d(\boldsymbol{x}, \boldsymbol{y})$	distance metric between two vectors
$\boldsymbol{c}$	centroids of codebook
Δ	pertubation of binary splitting
ϵ	treshold value
i	speech codebook entry
j	noise codebook entry
$H(\omega)$	filter function
$\boldsymbol{A}$	transition matrix of speech
$\boldsymbol{B}$	transition matrix of noise
$\boldsymbol{\Phi}$	transition matrix of augemented state-space system
$\boldsymbol{P}[n\|N]$	Kalman filter covariance matrix
$\boldsymbol{K}[n]$	Kalman gain
π_k	mixing coefficients of a GMM
$\boldsymbol{\mu}_k$	mean vectors of a GMM
$\boldsymbol{\Sigma}_k$	covariance matrix of a GMM
α, β	shape parameters of a Gamma distribution
γ_k	responsibilities of a GMM
λ_x	speech precision
λ_w	noise precision

Contents

1 Introduction

Additive noise is one of the most undesired effects in signal processing applications. In a typical speech application, the user is surrounded by many noise sources such as talking people in the background, passing cars, rain or manufactoring noise from building sites. All these noise sources have a high non-stationarity in common which needs to be adressed in noise reduction applications. One of the most challenging aspects of speech enhancement is to remove the undesired noise to serve high-quality speech in terms of speech quality and speech intelligibility. Modern devices such as smartphones and hearing aids benefit from state-of-the-art speech enhancement systems and can deliver high-quality speech and audio. Typical speech enhancement systems which rely on STFT-based methods give amplitude-only solutions which can reduce stationary and non-stationary noise to some degree but with undesired effects such as musical noise [1]. In the past years phase enhancement methods were proposed which were able to further improve the reconstructed speech in terms of speech quality and speech intelligibility. In stationary environments, these methods give superior results. In non-stationary environments at low SNR conditions they still suffer and therefore methods were proposed based on a priori information of speech stored in codebooks. Codebook-based approaches work in the AR domain which was first proposed by Lim and Oppenheim [2] for speech enhancement. Based on this work, auditory domain constraints were incorporated for improving the convergence behaviour in [3]. To improve immunity of AR parameters to additive white Gaussian noise, a higher-order statistics approach was proposed in [4] where a Wiener filter was applied and the spectral estimation of the AR model was obtained by using third-order cumulants.

In later studies [5] the model was formulated as a state-space system leading to a Kalman filter based speech enhancement system. Still, the estimation of the STP parameters was done using the Yule-Walker equations and in [6] EM-based algorithms were used to achieve better STP estimates. Nevertheless, at low SNR conditions and highly-nonstationary noise these methods suffer from poor performance and they do not use any prior knowledge about speech and noise. One of first work investigating codebook-based enhancement was proposed in [7] which used a codebook constraint Wiener filter. One major drawback was the initialization process, which was done using the noisy observation frame which gives very inaccurate results at low SNR scenarios. A major paper was published by Srinivasan et. al. [8] which proposed a codebook-driven STP estimator for speech enhancement. The a priori information was obtained through codebooks and a search algorithm computed the best match within all codebook combinations. The same authors extented this principle in [9] to a Bayesian framework incorporating intra and inter-frame dependencies. A more recent work proposed by Rosenkranz et. al. [10] using cepstral smoothing and delta codebooks to obtain more robust noise estimates.

This thesis presents a new approach and builds up on the previously described methods by incorporating GMM based prior training and implementing an iterative enahncement algorithm based on variational Bayesian approximation. The variational Bayesian algorithm is related to the conventional EM algorithm and derived as a special case if one seeks to find maximum-likelihood estimates of STP parameters.

The thesis is organized as follows; Chapter 2 gives a short introduction of variational Bayesian methods. Chapter 3 presents an overview of codebook based methods including ML and MMSE estimators of speech and noise STP parameters. Chapter 4 presents Kalman filter based speech

enhancement methods which rely on an EM-based parameter estimation framework. The proposed method is presented in Chapter 5 with detailed description of the used algorithms and results and performance evaluation are reported in Chapter 6. Chapter 7 concludes the work and gives a brief outlook on future topics and future work.

2 Background and Preliminaries

This chapter introduces the reader to basic concepts needed to understand the derived methods in this thesis. This chapter is organized as follows: Section 2.1 describes basic concepts of variational Bayesian methods and gives a detailed explanation of the VB-EM algorithm.

2.1 Variational Bayesian Inference

2.1.1 Motivation

In many estimation problems, one seeks to infer the joint *a posteriori* density $p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ of parameters $\boldsymbol{\theta}$ and hidden variables $\boldsymbol{x}$ given some noisy observation vector $\boldsymbol{y}$. This class of estimation problems is refered as Bayesian estimation, since the parameters and hidden variables are modeled as random variables. In signal processing applications the hidden states may comprise a desired signal, such as speech and an undesired portion like background noise whereas the parameters are some quantity of interest which are obtained from the observed signal, e.g. in noise reduction we seek to estimate the magnitude spectrum of clean speech. In range estimation we use time delay information to estimate the position of an obstacle. Robust estimators are essential for accurate parameter estimation at low SNR scenarios. The Bayesian formulation allows to incorporate prior knowledge based on prior densities and the posterior distribution is then obtained using Bayes formula

$$p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{x})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \tag{2.1}$$

where $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{x})$ is the likelihood, $p(\boldsymbol{\theta})$ is the prior density and $p(\boldsymbol{y})$ is called the evidence. In classical estimation theory, estimates are obtained as point estimates by computing ML, MAP or MMSE solutions [11]. Note, that these methods neglect the evidence in Bayes theorem and thus, only give an approximate solution. This simplification is often argued with the mathematically intractability of the evidence, i.e. resulting in high-dimensional integrals with no closed-form solutions. The intractability is even worse if the underlying models are not within a simple density family but rather have some arbritrary complex form. The MAP estimate is obtained as

$$\text{MAP}: \quad \hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{x})p(\boldsymbol{\theta}) \tag{2.2}$$

Deterministic approximation methods such as *variational Bayesian approximation* were developed over the past years which allow an approximation of the whole posterior distribution by a simpler, tractable distribution [12]. Contrary to Monte-Carlo methods such as MCMC [13] variational methods give an closed-form solution of the problem and are less complex regarding the computational burden but suffer from poorer estimation results. Nevertheless they found many applications in different fields of science such as machine learning [14], communications [15] or biomedical signal processing [16] with promising results.

2.1.2 Principle

The principle of variational inference relies solely on the work of Euler and Lagrange which investigated and developed so-called *calculus of variations*. Their goal was to compute derivations of so-called *functionals* which can be understood as a mapping from a function to a output value of the functional. Consider for example, the entropy of a random variable defined by Shannon et. al. [17]

$$H(p) = \int p(x)\log\, p(x)dx \tag{2.3}$$

The question which one may ask is: *How does the value of the entropy changes if the input function changes by an infinitesimal small amount?*. The answer leads to the so-called *functional derivative* developed by Feynman. et. al. [18] which turns out to be a constrained optimization problem. For the case of Bayesian inference, soppose a Bayesian model with given prior distributions. By applying the logarithm to Eq.(2.3), the logarithm of the evidence can be rewritten as

$$\log\, p(\boldsymbol{y}) = \mathcal{F}(q(\boldsymbol{x}), q(\boldsymbol{\theta}), \boldsymbol{y}) + \mathcal{D}_{KL}(q||p) \tag{2.4}$$

where $\mathcal{F}(q(\boldsymbol{x}), q(\boldsymbol{\theta}), \boldsymbol{y})$ is the free energy given by

$$\mathcal{F}(q(\boldsymbol{x}), q(\boldsymbol{\theta}), \boldsymbol{y}) = \int q(\boldsymbol{z}, \boldsymbol{\theta})\log \left(\frac{p(\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{z}, \boldsymbol{\theta})} \right) d\boldsymbol{z}d\boldsymbol{\theta} \tag{2.5}$$

and $\mathcal{D}_{KL}(q||p)$ is the Kullback-Leibler distance [19] between the joint pdf $p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta})$ and an auxiliary distribution $q(\boldsymbol{z}, \boldsymbol{\theta})$

$$\mathcal{D}_{KL}(q||p) = -\int q(\boldsymbol{z}, \boldsymbol{\theta})\log \left(\frac{p(\boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{y})}{q(\boldsymbol{z}, \boldsymbol{\theta})} \right) d\boldsymbol{z}d\boldsymbol{\theta} \tag{2.6}$$

The evidence can now be lower bounded by differentiating Eq.(2.6) with respect to $q(\boldsymbol{z}, \boldsymbol{\theta})$ or equivalently, minimizing the Kullback-Leibler divergence between the joint pdf and the auxiliary distribution. Now suppose that the we obtain a Kullback-Leibler divergence of zero. By inspection we observe that the auxiliary distribution $q(\boldsymbol{z}, \boldsymbol{\theta})$ would equal the posterior pdf $p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta})$ which is our desired result. In most practical situations, this is not case and only a lower bound is obtained. The goal of the variational inference engine is know to lower bound the evidence as good as possible - or in other words - to approximate the auxiliary distribution as close to the true posterior as possible. To do so, some assumptions and restrictions have to be introduced in order to guarante mathematically tractability. One assumption which has to be made is to restrict the auxiliary distribution to some known family of distributions. Another assumption is to allow factorization of the auxiliary distribution, thus assuming independence of the individual components.

2.1.3 Mean Field approach

The most simple way of restriction is to assume that the auxiliary distributions factorizes, i.e. the components are assumed to be independent

$$q(\boldsymbol{z}, \boldsymbol{\theta}) = q(\boldsymbol{z})q(\boldsymbol{\theta}) = \prod_{i=1}^{K} q_i(\boldsymbol{z}) \prod_{j=1}^{L} q_j(\boldsymbol{\theta}) \tag{2.7}$$

At a first glance, this assumption seems drastic since we decouple the hidden states and parameters which is not true in a real scenario. Nevertheless, it yields a very straightforward mathematical treatment and fairly good results. Also, the decoupling can be understood as replacing stochastic dependencies with deterministic dependencies between relevant moments of $\boldsymbol{z}$ and $\boldsymbol{\theta}$ as stated by Beal et. al. [20]. The above factorization is also known as *mean field approach* which was motivated from mean field theory of physics. Applying the mean-field approach to the evidence bound we obtain

$$\mathcal{F}(q(\boldsymbol{x}), q(\boldsymbol{\theta}), \boldsymbol{y}) = \int q_i(\boldsymbol{z}_i) q_j(\boldsymbol{\theta}_j) \log \tilde{p}(\boldsymbol{y}, \boldsymbol{\theta}_j, \boldsymbol{z}_i) d\boldsymbol{z}_i d\boldsymbol{\theta}_j - \int q_i(\boldsymbol{z}_i) \log\, q_i(\boldsymbol{z}_i) d\boldsymbol{z}_i \tag{2.8}$$

$$- \int q_j(\boldsymbol{\theta}_j) \log\, q_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j + const. \tag{2.9}$$

where the last two terms equals the entropy of $q(\boldsymbol{z}_i)$ and $q_j(\boldsymbol{\theta}_j)$, respectively. The distribution $\tilde{p}(\boldsymbol{y}, \boldsymbol{\theta}_j, \boldsymbol{z}_i)$ is obtained by computing the expectation of the joint pdf for all components $j \neq i$, i.e.

$$\log\, \tilde{p}(\boldsymbol{y}, \boldsymbol{\theta}_j, \boldsymbol{z}_i) = \mathbb{E}_{j \neq i} \Big\{ \log\, p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}) \Big\} + const. \tag{2.10}$$

Maximizing Eq.(2.10) yields the optimal distributions

$$\log\, \hat{q}_i(\boldsymbol{z}_i) = \mathbb{E}_{j \neq i} \Big\{ \log\, p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}) \Big\} \tag{2.11}$$

$$\log\, \hat{q}_j(\boldsymbol{\theta}_j) = \mathbb{E}_{k \neq j} \Big\{ \log\, p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}) \Big\} \tag{2.12}$$

2.1.4 Variational-Bayesian Expectation Maximization algorithm

Suppose we observe data $\boldsymbol{y}$ corrupted by noise. Let us denote the parameters of interest as $\boldsymbol{\theta}$ and hidden variables as $\boldsymbol{z}$. The evidence can then be lower bounded by

$$\log\, p(\boldsymbol{y}) \geq \int q(\boldsymbol{z}, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta})}{q(\boldsymbol{x}, \boldsymbol{\theta})} d\boldsymbol{x} d\boldsymbol{\theta} = \mathcal{F}(q(\boldsymbol{x}), q(\boldsymbol{\theta}), \boldsymbol{y}) \tag{2.13}$$

Applying the mean-field approach we obtain the auxiliary distributions as $q(\boldsymbol{z})q(\boldsymbol{\theta})$. Minimizing the lower bound in an iterative fashion leads to an EM-like algorithm which was first proposed by Beal. et. al. in [20]. The algorithm minimizes the lower bound by solving the auxiliary distributions in an round-robin fashion leading to two fundamental update equations

$$\textbf{VBE-Step}: \quad q^{(t+1)}(\boldsymbol{z}) \propto \exp \left(\int \log\, p(\boldsymbol{z}, \boldsymbol{y} | \boldsymbol{\theta}) q^{(t)}(\boldsymbol{\theta}) d\boldsymbol{\theta} \right) \tag{2.14}$$

$$\textbf{VBM-Step}: \quad q^{(t+1)}(\boldsymbol{\theta}) \propto \exp \left(\int \log\, p(\boldsymbol{z}, \boldsymbol{y} | \boldsymbol{\theta}) q^{(t)}(\boldsymbol{z}) d\boldsymbol{\theta} \right) p(\boldsymbol{\theta}) \tag{2.15}$$

Note the strong similarity to the standard EM algorithm. However, the VB-EM algorithm lower bounds the evidence of the model which gives an approximation of the true posterior rather then point estimates. A schematic overview of the principle is shown in Figure 2.1.

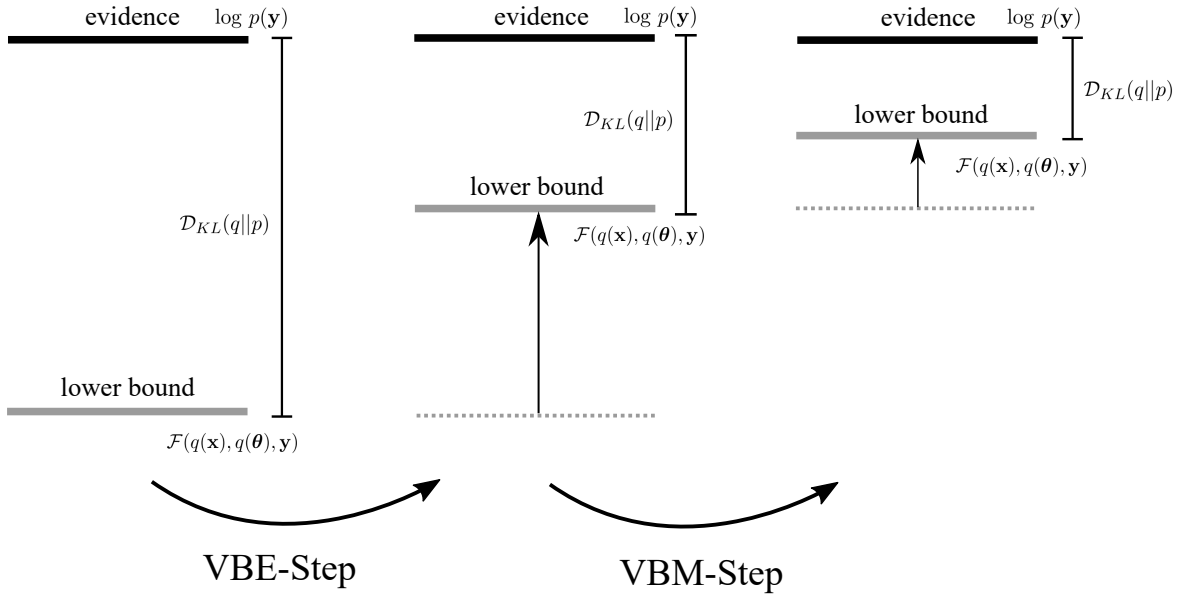

Figure 2.1: Principle of the VB-EM algorithm. The VBE-Step computes the variational distribution over hidden variables and the VBM-Step computes the variational distributions over parameters. The bound is tightened after every iteration or left unchanged and therefore the algorithm is guaranteed to converge.

The equivalent formulation of the standard EM algorithm would give

$$\textbf{E-Step}: \quad q^{(t+1)}(\boldsymbol{z}) = \underset{\boldsymbol{z}}{\operatorname{argmax}}\ \mathcal{F}(q(\boldsymbol{x}), q(\boldsymbol{\theta}^{(t)}), \boldsymbol{y}) \tag{2.16}$$

$$\textbf{M-Step}: \quad \boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \mathcal{F}(q(\boldsymbol{x}^{(t+1)}), q(\boldsymbol{\theta}), \boldsymbol{y}) \tag{2.17}$$

where the E-Step leads to the so-called *Q-function* given by

$$q^{(t+1)}(\boldsymbol{z}) = \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\theta}^{(t)}) \tag{2.18}$$

Furthermore, the VB-EM algorithm implies prior distributions of parameters which is not the case in the standard EM algorithm. The mean-field approach and further derivations are only valid if the models fulfill the conjungate-exponential conditions

1.) *The complete data likelihood is within the exponential family.*

2.) *The prior distribution is conjungate to the complete data likelihood.*

The VB-EM algorithm converges and obtains a local minima if and only if these two conditions are fulfilled [20]. Table 2.1 summarizes the VB-EM and the standard EM method for comparison.

Table 2.1: Comparison of standard EM and VB-EM algorithm.

EM-algorithm	**VB-EM algorithm**
E-Step: compute	**VBE-Step**: compute
$q^{(t+1)}(\boldsymbol{z}) = p(\boldsymbol{y}\|\boldsymbol{z}, \boldsymbol{\theta}^{(t)})$	$q^{(t+1)}(\boldsymbol{z}) \propto \exp\left(\int \log p(\boldsymbol{z}, \boldsymbol{y}\|\boldsymbol{\theta}) q^{(t)}(\boldsymbol{\theta}) d\boldsymbol{\theta}\right)$
M-Step: compute	**VBM-Step**: compute
$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ p(\boldsymbol{y}\|\boldsymbol{z}, \boldsymbol{\theta}^{(t)})$	$q^{(t+1)}(\boldsymbol{\theta}) \propto \exp\left(\int \log\ p(\boldsymbol{z}, \boldsymbol{y}\|\boldsymbol{\theta}) q^{(t)}(\boldsymbol{z}) d\boldsymbol{\theta}\right) p(\boldsymbol{\theta})$

3

Knowlegde-based Speech Enhancement

3.1 Motivation

Convential speech enhancement systems can operate well in stationary noise environments by using VAD-based [21] or minimum statistic based [22] noise PSD estimates. In non-stationary situations however, these methods provide inaccurate results and suffer from non-stationarity, especially at low SNR scenarios. Martin et. al. proposed a noise PSD estimator in [22] which employs a buffer of past samples producing long-term estimates of noise which gives promising results in stationary environments. As noise changes quickly, the performance is limited due to the buffer size and thus, performance is degraded. Kuropatwinksi et. al. [23] proposed a method using *codebooks* (CB) to estimate the short-term predictor (STP) parameters for speech coding applications. STP parameters are given by the linear predictive coefficients and excitation variance of the speech process. A speech enhancement framework was proposed by Srinivasan et. al. to overcome limitations in non-stationary environments by providing *a priori* information of speech and noise described by *short-term predictor* (STP) parameters in form of codebooks using maximum-likelihood solutions. The CB can be understood as a look-up table consisting pre-trained spectral shapes by applying vector quantization (VQ) techniques on training data. This work was further extended in [9] to a Bayesian estimation scheme were also inter-frame dependencies were investigated and integrated into the estimation process.
Promising results were reported in [10] dealing with highly non-stationary noise scenarios by incorporating *delta-codebooks* and using cepstral smoothing methods.

3.2 Signal Model

Noisy speech is modeled as an additive noise model where speech and noise are assumed to be independent

$$y[n] = x[n] + w[n] \tag{3.1}$$

where $y[n]$,$x[n]$ and $w[n]$ are the noisy speech, clean speech and noise and $n = 1, \ldots, N$, respectively. Modeling the clean speech and noise process as an autoregressive process yields a probability density of speech given the LP coefficients a

$$p(\boldsymbol{x}|\boldsymbol{a}_x) = \frac{1}{(2\pi)^{N/2} \det \boldsymbol{R}_x^{1/2}} \exp\Big(\frac{1}{2}\boldsymbol{x}^T \boldsymbol{R}_x^{-1} \boldsymbol{x}\Big) \tag{3.2}$$

where $\boldsymbol{a}_x$ is the vector of LP coefficients, N is the framelength in samples and $\boldsymbol{R}_x = \sigma_x^2 (\boldsymbol{A}_x^T \boldsymbol{A}_x)^{-1}$ is the autocorrelation matrix where $\boldsymbol{A}_x$ is a $N \times N$ lower triangular Toeplitz matrix and σ_x^2 is the excitation variance. The first column of $\boldsymbol{A}_x$ is given by $[1\ a_{x,1}\ \ldots a_{x,P}\ 0\ \ldots 0]$. The probability density of the noise process is defined analogously. The power spectrum of speech is given by

$$S_{xx}(\omega) = \frac{\sigma_x^2}{|1 + \sum_{p=1}^{P} a_{x,p} e^{-j\omega p}|^2} \tag{3.3}$$

from the inverse filter model [24]. The model order of speech is typical chosen in the interval $p \in [10\ 16]$ which leads to good performance. Higher predictor orders model higher formants in the envelope spectrum which introduces speaker dependence. To remove the speaker dependence, low model orders are chosen. For the noise, the model order depends on the noise type, e.g. babble noise has a speech-like characteristic and is therefore chosen in the same range as speech. Since the noise process is also modeled as an autoregressive model, the pdf of noise conditioned on the LP coefficients is equivalently obtained as

$$p(\boldsymbol{w}|\boldsymbol{a}_w) = \frac{1}{(2\pi)^{N/2} \det \boldsymbol{R}_w^{1/2}} \exp\Big(\frac{1}{2}\boldsymbol{w}^T \boldsymbol{R}_w^{-1} \boldsymbol{w}\Big) \tag{3.4}$$

where $\boldsymbol{a}_w$ and $\boldsymbol{R}_w$ are defined analogousely as speech. The noise power spectrum is given by

$$S_{ww}(\omega) = \frac{\sigma_w^2}{|1 + \sum_{q=1}^{Q} a_{w,q} e^{-j\omega q}|^2} \tag{3.5}$$

Before the prior information is obtained, the speech and noise spectra are gain-normalized to obtain the gain-normalized spectrum as

$$\overline{S}_{xx}(\omega) = \sigma_x^2 S_{ww}(\omega) \tag{3.6}$$
$$\overline{S}_{ww}(\omega) = \sigma_w^2 S_{ww}(\omega) \tag{3.7}$$

3.3 Codebook Training

To obtain the *a priori* information an appropriate training algorithm must be applied. Vector quantization methods were succesful used in speech coding, e.g. in [23] and described in [25] and applied in [8] for training of the AR parameters. Generating the codebook consists of three major steps

- Segmenting the training sequence into frames of 32ms duration
- Computing the AR coefficients by using standard methods like the autocorrelation method [24]
- Quantizing the AR coefficient vector using an appropriate algorithm

3.3.1 Linde-Buzo-Gray algorithm

A p-dimensional vector quantizer describes a mapping from a p dimensional vector to a finite subset $\mathcal{C}$ containing N codevectors. The Linde-Buzo-Gray algorithm [26] or often called the *generalized Lloyd algorithm* (GLA) is one of the most used algorithms for codebook generation and its principle is related to the k-means clustering approach which gains high popularity for data clustering in machine learning applications [12]]. The LBG algorithm is a generalization of the Lloyd algorithm developed by Stuart P. Lloyd [27]. The iterative procedure aims to find an optimal codebook of the input vector by applying the Lloyd iteration until convergence. If the statistics of the input vector are assumed to be known the Lloyd iteration is given by

Lloyd Iteration - Known Statistics

(*ii*) For a given codebook $\mathcal{C}_m = \{\boldsymbol{y}_i;\ i = 1, \ldots, N-1\}$ find the partition which forms

nearest neighbor cells

$$R_i = \{\boldsymbol{x} : d(\boldsymbol{x}, \boldsymbol{y}_i) < d(\boldsymbol{x}, \boldsymbol{y}_i) \forall j \neq i\}$$

(*ii*) Compute the centroid $\mathcal{C}_{m+1} = \{\text{cent}(R_i);\ i = 1, \ldots, N\}$

Step (*i*) of the Lloyd iteration can only improve or leave the encorder unchanged, thus the average distortion of the quantizer cannot increase. From this consequence it follows a very important lemma which states that every Lloyd iteration must reduce or leave the average distortion unchanged [28].

Linde-Buzo-Gray algorithm

(*i*) Chose as initial codebook containing N codewords

(*ii*) Partition the training data into N clusters where each cluster is represented by inital codebook. For partitioning a distortion measure is used such as the square distance or the Itakuro-Saito measure.

(*iii*) Compute an average distortion measure

$$D_j = \frac{1}{N} \sum_{n=1}^{N} \sum_{\boldsymbol{x}_i \in \mathcal{P}_{\boldsymbol{c}_n}} d(\boldsymbol{x}_i, \boldsymbol{c}_n)$$

(*iv*) For each cluster, recalculate the centroids

$$\boldsymbol{c}_n = \frac{1}{Y_n} \sum_{\boldsymbol{y}_i \in \mathcal{P}_{\boldsymbol{c}_n}} \boldsymbol{x}_i$$

(*v*) recalculate the distortion

$$D_{j+1} = \frac{1}{N} \sum_{n=1}^{N} \sum_{\boldsymbol{x}_i \in \mathcal{P}_{\boldsymbol{c}_n}} d(\boldsymbol{x}_i, \boldsymbol{c}_n)$$

(*vi*) If $\frac{D_j - D_{j+1}}{D_{j+1}} < \epsilon$ select the codeword $\boldsymbol{c}_n$ as the optimum, else go to Step (*ii*) and repeat after the stopping criterion is fulfilled

There are different methods to obtain an initial codebook for the LBG algorithm. The most popular method known as *binary splitting* computes an initial codevector as the average mean of the training sequence and splits the obtained codevector by adding a perturbation to the mean value.

$$\boldsymbol{c}_0 = \frac{1}{N} \sum_{\boldsymbol{x}_i \in \mathcal{X}} \boldsymbol{x}_i \tag{3.8}$$

$$\boldsymbol{c}_1 = (1 + \Delta)\boldsymbol{c}_0$$
$$\boldsymbol{c}_2 = (1 - \Delta)\boldsymbol{c}_0$$

where Δ is the perturbation. The new observed codevectors are then again split until the desired number of codevectors is achieved. The LBG algorithm is applied to the STP parameters of speech and noise to obtain a speech and noise codebook of size N_x and N_w, respectively.

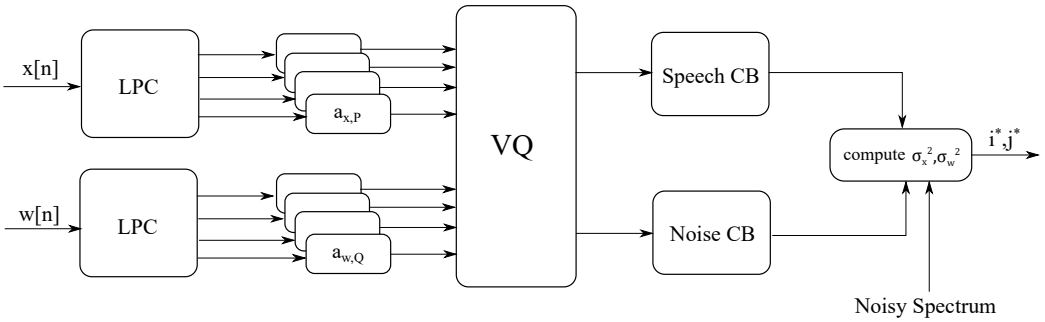

Figure 3.1: Overall training procedure of the CB-based approach.

3.4 Maximum Likelihood Estimation of STP Parameters

The estimates of clean speech and noise LP coefficients are given by the maximum-likelihood solution of the likelihood function, i.e. $\{i^*, j^*\} = \underset{i,j}{\operatorname{argmax}} \max_{\sigma_x^2, \sigma_w^2} p(\boldsymbol{y}|\boldsymbol{a}_x^i, \boldsymbol{a}_w^j, \sigma_x^2, \sigma_w^2)$. The likelihood function is a zero-mean Gaussian with covariance matrix $\boldsymbol{R}_y = \boldsymbol{R}_x + \boldsymbol{R}_w$.
Since the model order is small compared to the framelength one can approximate the Toeplitz matrices by circulant matrices [29]. This leads a description in the frequency domain, since the circulant matrices are diagonalized by the Fourier transform. Hence, we obtain the log-likelihood as

$$\mathcal{L} = \int_0^{2\pi} -\frac{S_{yy}(\omega)}{\frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2}} + \ln\left(\frac{1}{\frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2}}\right) d\omega \tag{3.9}$$

where $S_{yy}(\omega)$ is the noisy power spectrum, $A_x^i(\omega)$ is the spectrum of the i^{th} speech codebook entry and $A_w^j(\omega)$ is the spectrum of the j^{th} noise codebook entry, respectively. These spectra are given by taking the Fourier transform with respect to the LP coefficients

$$A_x^i(\omega) = \sum_{p=1}^{P} a_{x,p}^i e^{-j\omega p} \tag{3.10}$$

$$A_w^j(\omega) = \sum_{q=1}^{Q} a_{x,q}^j e^{-j\omega q} \tag{3.11}$$

It can be shown [30] that the log-likelihood function in Eq.(3.9) is equivalent as minimizing the Itakuro-Saito distance between the noisy power spectrum and the combined spectra for each codebook entry. The Itakuro-Saito distance is a measure of difference between two spectra and given by [24]

$$d_{IS}(S_{xx}(\omega), \hat{S}_{xx}(\omega) = \frac{1}{2\pi} \int_0^{2\pi} \left(\frac{S_{xx}(\omega)}{\hat{S}_{xx}(\omega)} - \ln\left(\frac{S_{xx}(\omega)}{\hat{S}_{xx}(\omega)}\right) - 1\right) d\omega \tag{3.12}$$

where $\hat{S}_{xx}(\omega)$ is an approximation of the noisy spectrum. The approximation can be constructed for each codebook entry of noise and speech

$$\hat{S}_{xx}^{i,j}(\omega) = \frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2} \tag{3.13}$$

Inserting Eq.(3.10) and (3.11) into Eq.(3.9) gives the optimum codebook indices as

$$\{i^*, j^*\} = \text{argmin}_{i,j}\left\{\min_{\sigma_w^2,\sigma_w^2} d_{IS}\left(S_{xx}(\omega), \frac{\sigma_x^2}{|A_x^i(\omega)|^2} + \frac{\sigma_w^2}{|A_w^j(\omega)|^2}\right)\right\} \tag{3.14}$$

The excitation variances are obtained by computing the derivative of the log-spectral distortion between the observed noisy spectrum and codebook combinations. This assumption is valid if the modeling error between $S_{xx}(\omega)$ and $\hat{S}_{xx}(\omega)$ is assumed to be small. Hence, the log-spectral distortion is obtained by using a Taylor series expansion up to seconds term of the logarithm in Eq.(3.9) [31]

$$d_{IS}(S_{xx}(\omega), \hat{S}_{xx}^{i,j}(\omega)) \approx \frac{1}{2} d_{LS}(S_{xx}(\omega), \hat{S}_{xx}^{i,j}(\omega)) \tag{3.15}$$

where the $d_{LS}(S_{xx}(\omega), \hat{S}_{xx}^{i,j}(\omega))$ is the log-spectral distortion between noisy observation and codebook entries. Differentiating Equation (3.13) with respect to the excitation variances leads the maximum likelihood solutions which can be written as a system of linear equations

$$\boldsymbol{C}\begin{bmatrix}\sigma_x^2\\ \sigma_w^2\end{bmatrix} = \boldsymbol{D} \tag{3.16}$$

with matrices

$$\boldsymbol{C} = \begin{bmatrix} \int_0^{2\pi} \left|\frac{1}{S_{yy}^2(\omega)|A_x^i(\omega)|^4}\right| d\omega & \int_0^{2\pi} \left|\frac{1}{S_{yy}^2(\omega)|A_x^i(\omega)|^2|A_w^j(\omega)|^2}\right| \\ \int_0^{2\pi} \left|\frac{1}{S_{yy}^2(\omega)|A_x^i(\omega)|^2|A_w^j(\omega)|^2}\right| & \int_0^{2\pi} \left|\frac{1}{S_{yy}^2(\omega)|A_w^j(\omega)|^4}\right| d\omega \end{bmatrix}$$

$$\boldsymbol{D} = \begin{bmatrix} \int_0^{2\pi} \left|\frac{1}{S_{yy}(\omega)|A_x^i(\omega)|^2}\right| d\omega \\ \int_0^{2\pi} \left|\frac{1}{S_{yy}(\omega)|A_w^j(\omega)|^2}\right| d\omega \end{bmatrix}$$

The excitation variances can obtained by solving Eq.(3.13)

$$\begin{bmatrix}\sigma_x^2\\ \sigma_w^2\end{bmatrix} = \boldsymbol{C}^{-1}\boldsymbol{D} \tag{3.17}$$

Excitation variances which lead negative estimates are excluded from the set. Figure 3.2 shows a block diagram of the overall estimation scheme. Note that these procedure is done for every codebook combination which can be interpreted as a full search algorithm until the candidate pair with the lowest Itakuro-Saito distance is found.

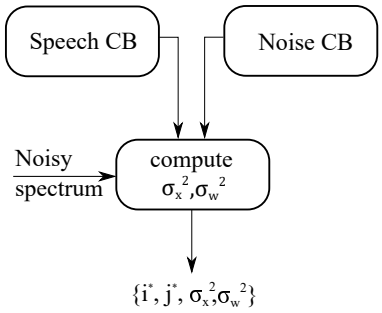

Figure 3.2: Estimation procedure of the codebook-based approach.

The complete estimation strategy of the CB approach is summarized in Algorithm 1.

for *every frame* **do**
  set $D = \infty$;
  for $i = 1 : N_x$ **do**
    for $j = 1 : N_w$ **do**
      $\hat{\sigma}_x^2, \hat{\sigma}_w^2 = \text{argmin}_{\sigma_x^2, \sigma_w^2} d_{IS}(S_{xx}(\omega), \hat{S}_{xx}^{i,j}(\omega))$;
      if $\sigma_x^2, \sigma_w^2 \geq 0$ **then**
        if $d_{IS}(S_{xx}(\omega), \hat{S}_{xx}^{i,j}(\omega)) < D$ **then**
          $\{i^*, j^*, \hat{\sigma}_x^2, \hat{\sigma}_w^2\} = \{i, j, \sigma_x^2, \sigma_w^2\}$;
          $D = d_{IS}(S_{xx}(\omega), \hat{S}_{xx}^{i,j}(\omega))$;
        end
      end
    end
  end
end

Algorithm 1: ML estimation of STP parameters using speech and noise codebook *a priori* information [8].

After the optimum STP parameters were found, the enhanced speech is obtained by applying a convential Wiener filter which is constructed from the estimated STP parameters

$$H(\omega) = \frac{\frac{\hat{\sigma}_x^2}{|A_x^i(\omega)|^2}}{\frac{\hat{\sigma}_x^2}{|A_x^i(\omega)|^2} + \frac{\hat{\sigma}_w^2}{|A_w^j(\omega)|^2}} \tag{3.18}$$

The so derived Wiener filter is applied to the noisy DFT coefficients to reconstruct the enhanced speech signal.

4

Kalman-Filter based Speech Enhancement

This chapter introduces the reader to the Kalman-Filter based Speech Enhancement method first proposed in by Paliwal et. al. [32] and further developed as an iterative and sequential speech enhancement system by Gannot et. al. [6] which is based on the Expectation-Maximization (EM) algorithm. This chapter is organized as follows: Section 4.1. points out the motivation and background for Kalman-based speech enhancement. Section 4.2 introduces the signal model and state-space description of the whole system. Finally, Section 4.3 discusses the estimation procedure and the EM algorithm in combination with the Kalman filter.

4.1 Motivation

The Kalman filter in speech enhancement applications was first proposed by Paliwal et. al. in [32] where experiments showed superior performance over the Wiener filter if the clean speech parameter are assumed known. The above mentioned method was only investigated for white noise and so the authors in [33] presented a model where colored noise types where incorporated in the state-space model leading better results for a broad type of noise sources. Nevertheless, none of these methods describes the parameter estimation procedure, a crucial aspect since the Kalman filter only gives acceptable results if vigorous parameter estimates are available. Gannot et. al. presented in [6] an EM-based approach of the parameter estimation problem in combination with the Kalman filter methods from earlier papers. Motivated from these works further EM-based approaches like in [34] or [35] using a harmonic speech model were presented each yielding superior performance for a vast tpye of noise scenarios. The principles presented by Gannot et. al. are used and further extended in the proposed method presented in Chapter 5 and are therefore essential for this thesis.

4.2 Signal Model and System Representation

4.2.1 Speech Model

Clean speech $x[n]$ is assumed to be modeled as an AR(p)-model with zero-mean white Gaussian innovation process $u[n]$ with precision λ_x. The speech sample at time instant n can be formulated as

$$x[n] = -\sum_{k=1}^{P} a_{x,k} x[n-k] + u[n] \tag{4.1}$$

where $a_{x,k}$ is the kth linear prediction coefficient and P is the model order. In vector notation, Eq.(4.1) reads

$$x[n] = -\boldsymbol{a}_x^T \boldsymbol{x}_p[n-1] + u[n] \tag{4.2}$$

where $\boldsymbol{a}_x^T = [a_{x,1}\ a_{x,2}\ \dots\ a_{x,P}]$ and $\boldsymbol{x}_p[n-1] = [x[n-1]\ x[n-2]\ \dots\ x[n-P]]^T$ are the P past samples of clean speech. Since the innovation process is Gaussian, the parametric distribution

of the clean speech is given as

$$x[n] \sim \mathcal{N}(\boldsymbol{a}_x^T \boldsymbol{x}_p[n-1], \mathbf{I}\sigma_x^2)$$

4.2.2 Noise Model

The noise process $w[n]$ is modeled as an AR(q)-process with zero-mean white Gaussian innovation. Note that, in general, the clean speech order and noise order is not equal. The noise sample at time instant n can be formulated as

$$w[n] = -\sum_{k=1}^{Q} a_{w,k} v[n-k] + \nu[n] \tag{4.3}$$

where $a_{w,k}$ is the kth linear prediction coefficient and Q is the model order. In vector notation, Eq.(4.3) reads

$$w[n] = -\boldsymbol{a}_w^T \boldsymbol{w}_q[n-1] + \nu[n] \tag{4.4}$$

where $\boldsymbol{a}_w^T = [a_{w,1}\ a_{w,2}\ \ldots\ a_{w,Q}]$ and $\boldsymbol{w}_q[n-1] = [w[n-1]\ w[n-2]\ \ldots\ w[n-Q]]^T$ are the Q past samples of the noise. Since the innovation process is Gaussian, the parametric distribution of the noise is given as

$$w[n] \sim \mathcal{N}(\boldsymbol{a}_w^T \boldsymbol{w}_q[n-1], \lambda_w^{-1})$$

4.2.3 State-Space Model

Speech State-Space Model

The state-space model of clean speech comprises the additive observation model and the AR source model of order P which are given by

$$\text{State equation :} \quad \boldsymbol{x}[n] = \boldsymbol{A}\boldsymbol{x}[n-1] + \boldsymbol{u}[n] \tag{4.5}$$

$$\text{Observation equation :} \quad \boldsymbol{y}[n] = \boldsymbol{x}[n] + \boldsymbol{w}[n] \tag{4.6}$$

where $\boldsymbol{A}$ is the $P \times P$ state transition matrix given by

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & \ldots & \ldots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & \ldots & 0 & 1 \\ -a_{x,p} & -a_{x,p-1} & \ldots & \ldots & -a_{x,2} & -a_{x,1} \end{bmatrix}$$

and $\boldsymbol{u}[n]$ is the innovation process which is modeled as a zero-mean Gaussian distribution with variance σ_x^2. $\boldsymbol{x}[n-1]$ is a vector containing the P past samples of $x[n]$.

Noise State-Space Model

The state-space formulation of the noise process is analog to the speech case, hence we have

$$\text{State equation :} \quad \boldsymbol{w}[n] = \boldsymbol{B}\boldsymbol{w}[n-1] + \boldsymbol{\nu}[n] \tag{4.7}$$

$$\text{Observation equation}: \quad w[n] = \boldsymbol{w}[n]\boldsymbol{h}_w^T \tag{4.8}$$

where $\boldsymbol{B}$ is the $Q \times Q$ state transition matrix given by

$$\boldsymbol{B} = \begin{bmatrix} 0 & 1 & 0 & \dots & \dots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & 0 & 1 \\ -a_{w,q} & -a_{w,q-1} & \dots & \dots & -a_{w,2} & -a_{w,1} \end{bmatrix}$$

and $\boldsymbol{\nu}[n]$ is the innovation process of noise.

Augmented State-Space Model

For colored noise types and non-stationary noise types, the complete state-space model of the observed speech is obtained by augmenting Eq.(4.5)-Eq.(4.8) which yields a state-space model for the complete data set also known as *perfect measurement representation*

$$\begin{aligned} &\text{State Equation}: \quad \boldsymbol{z}[n] = \boldsymbol{\Phi}\boldsymbol{z}[n-1] + \boldsymbol{G}\boldsymbol{r}[n] \\ &\text{Observation equation}: \quad y[n] = \boldsymbol{h}^T\boldsymbol{z}[n] \end{aligned} \tag{4.9}$$

where $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{bmatrix}$ is the state transition matrix of the complete system, $\mathbf{r}[n] = [u[n]\ \nu[n]]$ is a vector of innovation processes and $\mathbf{G} = \begin{bmatrix} \sigma_x & 0 \\ 0 & \sigma_w \end{bmatrix}$ contains the excitation variances, respectively.

4.3 Parameter Estimation

In order to enhance speech, knowledge of the clean STP parameters is needed. Since we only obtain a corrupted version a direct computation is not possible and would lead to degraded speech performance. Therefore, the authors proposed an EM-based framework for the estimation of parameters and hidden variables. As a side product, the EM algorithm gives signal estimates and decouples clean speech and noise estimation which leads to very elegant solutions.

4.3.1 EM-based STP estimation

The goal of the EM-based estimation approach is to estimate the clean speech $x[n]$ and its STP parameters from the observed noisy sequence $y[n]$. Since the speech and noise processes are not obtained individually by the measurement system, one can model the collection as a hidden variable $\boldsymbol{z} = [\boldsymbol{x}\ \boldsymbol{w}]$. In signal processing literature this is often called *complete data* which is related to the observed vector through a one-to-many transformation

$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{z}) \tag{4.10}$$

Applying the standard EM-algorithm [36] consits of two steps, namely the *expectation step* (E-Step) and *maximization step* (M-Step) which are given by

$$\text{E-Step}: \quad \mathcal{Q}\Big(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(t)}\Big) = \mathbb{E}_{\hat{\boldsymbol{\theta}}^{(t)}}\{\log\, p(\boldsymbol{z}; \boldsymbol{\theta}|\boldsymbol{y})\} \tag{4.11}$$

$$\text{M-Step}: \quad \hat{\boldsymbol{\theta}}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}} \mathcal{Q}\Big(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(t)}\Big) \tag{4.12}$$

where $\hat{\boldsymbol{\theta}}^{(t)}$ is the obtained estimate after t iterations. The so-called *Q-function* in the E-Step gives the *a posteriori* distribution of the complete data set, hence the signal estimates of speech and noise. The M-Step yields the maximum-likelihood solution of the estimation problem by differentiating the *a posteriori* density of the complete data. It is important to note that the E-Step consists of the Kalman filter recursion and that the M-Step yields equations similar to the well-known Yule-Walker equations.

E-Step

The goal in the E-Step is to compute the state estimate $\boldsymbol{z}[n|N]$ and the associated error covariance matrix $\boldsymbol{P}[n|N]$ based on the observation frame $y[n]$ where $n = 0, 1, \ldots, N-1$. The forward filtering recursion is the well-known Kalman filter [37] given by four main equations which are classified into *Propagation Equations* and *Update Equations*. The equations are given by:

Propagation Equations:

$$\boldsymbol{z}[n|n-1] = \tilde{\boldsymbol{\Phi}}\boldsymbol{z}[n-1|n-1] \tag{4.13}$$

$$\boldsymbol{P}[n|n-1] = \tilde{\boldsymbol{\Phi}}\boldsymbol{P}[n-1|n-1]\tilde{\boldsymbol{\Phi}}^T + \boldsymbol{G}\boldsymbol{G}^T \tag{4.14}$$

Update Equations:

$$\boldsymbol{z}[n|n] = \boldsymbol{z}[n|n-1] + \boldsymbol{K}[n]\Big(y[n] - \boldsymbol{h}^T\boldsymbol{z}[n|n-1]\Big) \tag{4.15}$$

$$\boldsymbol{P}[n|n] = \boldsymbol{P}[n|n-1] - \boldsymbol{K}^T[n]\boldsymbol{h}^T\boldsymbol{P}[n|n-1] \tag{4.16}$$

where $\boldsymbol{z}[n|n-1]$ and $\boldsymbol{P}[n|n-1]$ are also known as the *predicted* state vector and error covariance matrix at time step n which are computed *a priori* the observation. Note that the transition matrix already contains estimates of the parameters which are given from the previous M-Step. After a new measurement is observed, the update equations compute the *a posteriori* state vector and covariance matrix $\boldsymbol{z}[n|n]$ and $\boldsymbol{P}[n|n]$, respectively. $\boldsymbol{K}[n]$ is known in the literature as the *Kalman Gain* and is obtained as

$$\boldsymbol{K}[n] = \frac{\boldsymbol{P}[n|n-1]\boldsymbol{h}}{\boldsymbol{h}^T\boldsymbol{P}[n|n-1]\boldsymbol{h}} \tag{4.17}$$

After the *a posteriori* estimates up to time instant N are computed in the first stage, the smoothed estimates are computed in a backward recursion described by two fundamental equations

Smoothing recursion:

$$\boldsymbol{z}[n|N] = \boldsymbol{z}[n-1|n-1] + \boldsymbol{S}[n-1]\Big(\boldsymbol{z}[n-1|N] - \boldsymbol{\Phi}\boldsymbol{z}[n-1|n-1]\Big) \tag{4.18}$$

$$\boldsymbol{P}[n-1|N] = \boldsymbol{P}[n-1|n-1] - \boldsymbol{S}[n-1]\Big(\boldsymbol{P}(n|N) - \boldsymbol{P}[n|n-1]\Big)\boldsymbol{S}^T[n-1] \tag{4.19}$$

where $\boldsymbol{S}[n-1] = \boldsymbol{P}[n-1|n-1]\hat{\boldsymbol{\Phi}}^T\boldsymbol{P}^{-1}[n|n-1]$. It turns out that smoothed estimates of the signal yields a better representation and results in better speech quality. The recursion are applied sample-by-sample for each frame of the noisy speech.

M-Step

Given the signal model of Section 4.2 yields the complete data log-likelihood as

$$\log p(\boldsymbol{z};\boldsymbol{\theta}) = \log p(\boldsymbol{y};\boldsymbol{\theta}) + \log p(\boldsymbol{w};\boldsymbol{\theta}) \tag{4.20}$$

Inserting the Gaussian densities of the innovation processes in Eq.(4.20) yields

$$\log p(\boldsymbol{z};\boldsymbol{\theta}) = const. + \log p(\boldsymbol{x}_p(0)) + \log p(\boldsymbol{w}_q(0)) - \frac{N}{2}\log \sigma_x^2 - \frac{N}{2}\log \sigma_w^2 \tag{4.21}$$

$$- \frac{1}{2\sigma_x^2}\sum_{n=0}^{N-1}\Big(x[n] + \boldsymbol{a}_x^T\boldsymbol{x}_p[n-1]\Big)^2 - \frac{1}{2\sigma_w^2}\sum_{n=0}^{N-1}\Big(w[n] + \boldsymbol{a}_w^T\boldsymbol{w}_q[n-1]\Big)^2 \tag{4.22}$$

Note that the terms $\log p(\boldsymbol{x}_p(0))$ and $\log p(\boldsymbol{w}_q(0))$ can be neglected since the model order is much smaller than the framelength, wich is equivalent to ciruclant approximation of the Topelitz matrices of the Gaussian density. Computing the M-Step needs knowledge of the signal estimates which can be obtained by the Kalman filter recursions in the E-Step and yields the *Q-function* as

$$\mathcal{Q}\Big(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(t)}\Big) = const. - \frac{N}{2}\log \sigma_x^2 - \frac{N}{2}\log \sigma_w^2 \tag{4.23}$$

$$- \frac{1}{2\sigma_x^2}\sum_{n=0}^{N-1}\left(\widehat{x^2[n]} + 2\mathbf{a}_x^T\widehat{\mathbf{x}_p[n-1]x[n]} + \mathbf{a}_x^T\widehat{\mathbf{x}_p[n-1]\mathbf{x}_p^T[n-1]}\,\mathbf{a}_x\right) \tag{4.24}$$

$$- \frac{1}{2\sigma_w^2}\left(\widehat{w^2[n]} + 2\mathbf{a}_w^T\widehat{\mathbf{w}_q[n-1]w[n]} + \mathbf{a}_w^T\widehat{\mathbf{w}_q[n-1]\mathbf{w}_q^T[n-1]}\mathbf{a}_w\right) \tag{4.25}$$

The ML solutions are obtained by differentiating Eq.(4.23) with respect to the parameter vector $\boldsymbol{\theta} = [\boldsymbol{a}_x\ \boldsymbol{a}_w\ \sigma_x^2\ \sigma_w^2]$ which yields estimates of speech and noise STP parameters. The signal estimates are obtained from the state covariance matrix of the Kalman filter which is given as

$$\boldsymbol{P}(n|N) = \widehat{\boldsymbol{z}[n]\boldsymbol{z}^T[n]} - \widehat{\boldsymbol{z}[n]}\,\widehat{\boldsymbol{z}^T[n]} \tag{4.26}$$

and therefore are extracted as block matrices, i.e. $\widehat{\boldsymbol{x}[n]\boldsymbol{x}^T[n]}$ is the upper left $P \times P$ matrix of $\widehat{\boldsymbol{z}[n]\boldsymbol{z}^T[n]}$ and $\widehat{\boldsymbol{w}[n]\boldsymbol{w}^T[n]}$ is the lower right $Q \times Q$ matrix of $\widehat{\boldsymbol{z}[n]\boldsymbol{z}^T[n]}$, respectively.

Speech LP coefficients

The speech LP coefficients are obtained by differentiating Eq.(4.23) as

$$\hat{\mathbf{a}}_x = -\left(\sum_{n=0}^{N-1}\widehat{\mathbf{x}_p[n-1]\mathbf{x}_p^T[n-1]}\right)^{-1}\sum_{n=0}^{N-1}\widehat{\mathbf{x}_p[n-1]x[n]} \tag{4.27}$$

Speech excitation variance

The speech excitation variance is given by

$$\sigma_x^2 = \frac{1}{N}\sum_{n=0}^{N-1}\left(x^2[n] + \hat{\mathbf{a}}_x^T\widehat{\mathbf{x}_p[n-1]x[n]}\right) \tag{4.28}$$

Noise LP coefficients

The noise LP coefficients are obtained by differentiating Eq.(4.23) as

$$\hat{\mathbf{a}}_w = -\left(\sum_{n=0}^{N-1} \widehat{\mathbf{w}_q[n-1]\mathbf{w}_q^T[n-1]} \right)^{-1} \sum_{n=0}^{N-1} \mathbf{w}_q[n-1]w[n] \tag{4.29}$$

Noise excitation variance

The noise excitation variance is given by

$$\sigma_w^2 = \frac{1}{N} \sum_{n=0}^{N-1} \left(w^2[n] + \hat{\mathbf{a}}_w^T \widehat{\mathbf{w}_q[n-1]w[n]} \right) \tag{4.30}$$

The signal estimates are obtained from the covariance matrix of the Kalman filter recursion. Note that they have a very similar form as the Yule-Walker equations. The EM algorithm extracts the signal statistics from the Kalman filter contrary to the standard Yule-Walker equations. Since the approach is based on an iterative EM algorithm, it is guaranteed to converge and the log-likelihood function can be used as a treshold value for stopping criterion. The method suffered from severe degradation at low SNR scenarios due to the poor intialization which was done using the noisy observation frame. For voiced and unvoiced frames the Kalman filter is not able to compute the hidden states posterior distributions properly using the noisy initialization. Therefore, an initalization procedure based on higher-order statistics was proposed to circumvent this issue by using third-order cumulants. This initialization procedure is only valid for Gaussian noise scenarios where high-order cumulants of noise vanish [38]. The benefit lies in an significant increase of speech quality but the noise model is assumed to be additive white noise. The Kalman filter suffers mainly from the initialization procedure and thus a good and robust initialization needs to be implemented. The noise statistics could be either extracted from silent regions using VAD-based approaches or by using a noise PSD estimator. The speech statistics cannot directly be evaluated, most approaches use the noisy observation as initialization. This is a valid assumption for high SNR regions since speech is more present but the initialization fails at low SNR regions were noise masks speech.

5 Proposed Method

The main novelty in this thesis is based upon a full Bayesian estimation of AR parameters and excitation variances by using *a priori* knowledge in terms of pre-trained GMMs. To reach this goal, variational approximation methods are used, which are very young in the field of machine learning. Also, a state-space description combined with a Kalman filter is used, to develop an iterative framework which will be called the *Variational Bayesian Expectation-Maximization* (VB-EM) algorithm. Since the EM algorithm of the coventional Kalman-based method provides ML solutions only, it is of big interest to further develop a Bayesian estimation framework. Work was done using pre-trained speech and noise codebooks for deriving MMSE estimators of short-term predictive coefficients. A variational Bayesian approach for multichannel room dereverberation was proposed in [39]. This chapter is organized as follows: Section 5.1 describes the motivation of the new approach. Section 5.2. elaborates in detail the GMM training stage. Section 5.3 explains the variational Bayesian engine. Section 5.4 gives a summary of the algorithm and Section 5.5 concludes the chapter with a discussion and comparison to reference methods.

5.1 Motivation

For a practical usage of the obtained state-space system, the clean and noise STP parameters need to be estimated in order to enhance the speech signal. In Chapter 4, an EM-based framework was described, which estimates the STP parameters based on the signal estimates from a Kalman filter. No *a priori* knowlegde of the parameters was incorporated which leads to a maximum-likelihood solution similar to the well-known Yule-Walker equations. First attempts to incorporate *a priori* knowledge of STP parameters was obtained by using a codebook constrained iterative Wiener filter proposed by Sreenivas et. al. [7]. Codebook based methods were further investigated by Kuropatwinksi et. al. [23] and Srinivasan et. al. [8]. The *a priori* knowledge of AR parameters was obtained from pre-trained codebooks obtained from training data using vector quantization techniques such as the Linde-Buzo-Gray algorithm (LBG). To obtain the enhanced speech signal, a conventional Wiener filter was used. Since vector quantization uses a k-means type clustering algorithm, it provides centroids and associates each data point to a Voronoi cell by minimizing a distance metric. Gaussian Mixture Models on the other hand provide not only the centroids, but also the associated covariance matrices and weighting coefficients. Thus, GMM-based clustering can be understood as a soft-clustering approach which yields a full probabilistic model contrary to the hard-clustering approach of vector quantization. Therefore, GMMs for both speech and noise are trained *a priori* to obtain distributions of LP coefficients which reduce the complexity of the involved enhancement procedure and yield more accurate estimates. A variational Bayesian approach is implemented and investigated to obtain posterior distributions for the given probability model.

5.2 Prior model using GMMs

To obtain the *a priori* knowledge for the proposed method, *Gaussian Mixture Models* (GMMs) are trained on clean speech and noise. The prior distributions allow a complete probabilistic treatment of the estimation procedure and groups the training dataset into soft clusters by describing the uncertainty of the centroids in form of computed covariance matrices and responsibilities. GMMs were succesfully used in numerous applications such as database retrieval in image processing [40] or text-independent speaker identification [41]. They were also used for speech enhancement in earlier literature such as in [42] or [43]. Their most vigorous property is the ability to approximate any continuous density to arbritrary complexity which makes them essential in clustering the multimodal and multidimensional distributions [12]. They can be described mathematically as a linear combination of K Gaussians

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{5.1}$$

where π_k is the k^{th} mixing coefficient and $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ are the k^{th} mean vector and covariance matrix, respectively. Since the probability axioms must hold [44], the mixing coefficients must satisfy $0 \leq \pi_k \leq 1$ and must sum up to one

$$\sum_{k=1}^{K} \pi_k = 1 \tag{5.2}$$

It is useful for mathematical reasons, to introduce a binary indicator variable $\boldsymbol{\eta}$ which indicates the activation state of a certain GMM component with property $\sum_k \eta_k = 1$ which gives a marginal distribution $p(\boldsymbol{\eta})$ specified as

$$p(\eta_k = 1) = \pi_k \tag{5.3}$$

Thus, the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{\eta}$ is obtained as a Gaussian

$$p(\boldsymbol{x}|\eta_k = 1) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{5.4}$$

The introduction of the indicator variable leads to a simplified treatment and more tractable mathematical derivations. To obtain a GMM from an arbritrary data set, training algorithms must be applied to obtain the clusters.

5.2.1 EM-based GMM clustering

The most powerful method for GMM training is the *Expectation-Maximization* (EM) algorithm proposed by Dempster et. al. in [36] for finding maximum-likelihood solutions of the mean vectors and covariance matrices. The algorithm starts the initialization either with randomely generated centroid and covariance matrices or by user-defined parameters.

EM-algorithm for GMM Training

(i) Initialization of mean vectors $\boldsymbol{\mu}_k$, covariance matrices $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k.

(ii) **E-Step:** Compute the responsibilities

$$\gamma(z_k) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

(*iii*) M-Step: Compute the parameters using responsibilities from the E-Step

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{N}\sum_{n=0}^{N-1}\gamma(z_k)\boldsymbol{x}[n]$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{N}\sum_{n=0}^{N-1}\gamma(z_k)(\boldsymbol{x}[n]-\boldsymbol{\mu}_k^{(t+1)})^T(\boldsymbol{x}[n]-\boldsymbol{\mu}_k^{(t+1)})$$

To demonstrate the importance of GMMs, suppose a two-dimensional data set given in Figure 5.1, where the data forms three significant groups. A single Gaussian distribution obtained from the maximum-likelihood solution is not able to capture the whole strucutre of the data set and gives a very imprecise description. Gaussian Mixture Models are able to capture the multimodal distribution of the data set by applying the EM-algorithm. Using an appropriate number of mixture components, GMMs can approximate almost any continuous density to arbritrary complexity.

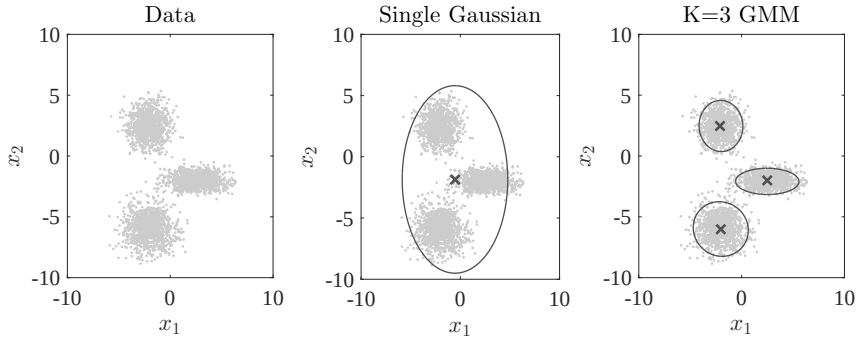

Figure 5.1: Example of a two-dimensional data set with 3 dominant clusters (left). The single Gaussian description (middle) cannot recover the whole structure, whereas the GMM (right) can describe complex densities.

Singularity Problem

A problem which may occur training GMMs is the so-called *singularity problem* which occurs when a data point collapses with one of the centroids. Suppose a GMM where a data point collapses with one of the centroids, i.e. the likelihood function has the form

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{x}, \sigma_j^2\boldsymbol{I}) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \tag{5.5}$$

which tends to infinity for $\sigma_j^2 \to 0$. This issue can be avoided by adding a regularization term to the covariance matrices or by resetting the covariance matrices and recompute the centroids.

5.2.2 Setup

The training data consisted of 10 minutes of speech from the TIMIT database [cite] from randomely chosen utterances. Before training, the data set was windowed using a Hamming window with length $N = 512$ samples and 50% overlap. For every frame the LP coefficients were computed using the standard Yule-Walker equations and converted to line-spectral frequencies. This step is crucial since quantizing line-spectral frequencies leads to stable filter coefficients due to

their smaller sensitivity to quantization noise which may not be achieved using LP coefficients [25]. The spectral envelopes of clean speech and noise were gain normalized to obtain a normalized spectrum

$$S_{xx}(\omega) = \sigma_x^2 \overline{S}_{xx}(\omega) \tag{5.6}$$

$$S_{ww}(\omega) = \sigma_w^2 \overline{S}_{ww}(\omega) \tag{5.7}$$

where $S(\omega)_{xx}$ is the clean speech power spectrum, $\overline{S}(\omega)_{xx}$ its normalized version and σ_x^2 represents a gain factor. The same notation holds for the noise process in further sections.
A codebook was obtained by applying the LBG algorithm to the LSF data set. The codebook was used for the initialization of the EM-algorithm since LBG initialization works more accurately than randomization and gives good initial cluster regions. Figure 5.2 exemplifies the difference between VQ training and the new GMM-based approach of babble noise data. For visualization, the first two dimension of the data set are depicted. The VQ clustering approach assigns each data point to a Voronoi cell which are bounded by the decision boundary. The GMM method refines the centroids from the VQ approach and computes associated covariance matrices with weighting factors leading to overlapping regions. This allows soft-clustering since every mixture component is weighted and describes its responsibility to the overall mixture model. Different codebook sizes and number of mixture components were trained and are further discussed in Chapter 6. Also, a self-regularization was used in the EM training, since many mixture components may lead to a singularity problem where the computed centroid collapses with a data point.

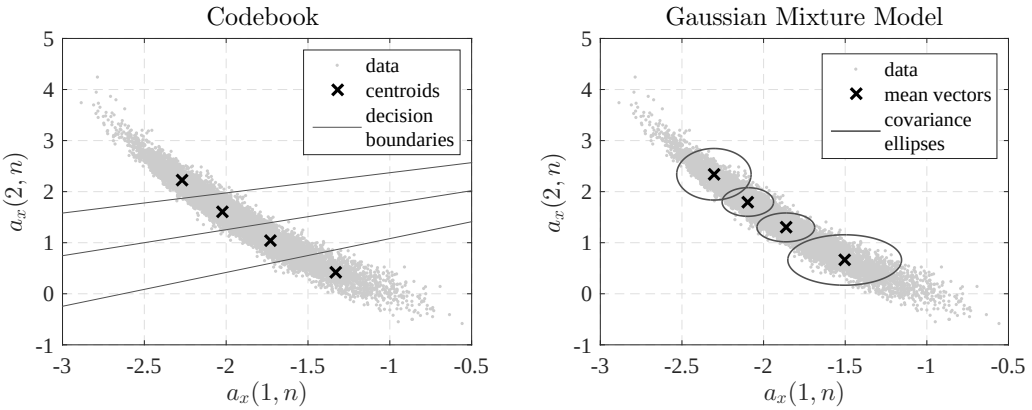

Figure 5.2: Vector Quantization training using LBG algorithm (left) and GMM-based EM training using the same data set of babble noise (right).

The overall training setup summarized in Table 5.1.

Table 5.1: Training setup.

Parameter	**Settings**
Sampling frequency f_s	16kHz
Frame length N	512
Window type	Hamming
Frame Overlap	50%

After training, the prior distributions of speech and noise coefficients are given by

$$p(\boldsymbol{a}_x) = \sum_{k=1}^{N_x} \pi_{x,k} \mathcal{N}(\boldsymbol{a}_x | \boldsymbol{\mu}_{a_x,k}, \boldsymbol{\Sigma}_{a_x,k}) \tag{5.8}$$

$$p(\boldsymbol{a}_w) = \sum_{k=1}^{N_w} \pi_{w,k} \mathcal{N}(\boldsymbol{a}_w | \boldsymbol{\mu}_{a_w,k}, \boldsymbol{\Sigma}_{a_w,k}) \tag{5.9}$$

where N_x and N_w are the size of the speech and noise GMM, respectively.

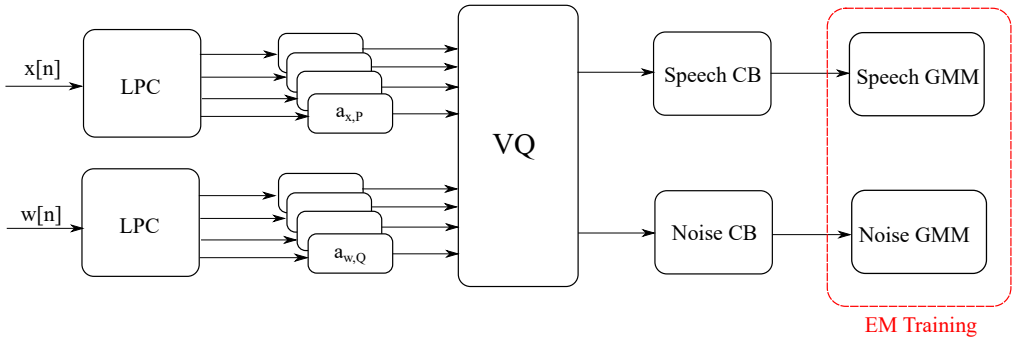

Figure 5.3: Block diagram of the GMM-based training stage.

Figure 5.3 shows a block diagram of the proposed training stage, including the CB-based pre-training and the EM-based GMM training.

5.3 Variational-Bayesian Parameter Estimation

The parameter estimation procedure is motivated from the methods described in Chapter 3 and 4. Thus, an EM-based algorithm is implemented using techniques from variational Bayesian approximation which enables the treatment with mixture model priors.

5.3.1 Probabilistic Model

The observation is modeled as an additive noise model where the noise and clean speech are assumed independent.

$$y[n] = x[n] + w[n] \tag{5.10}$$

Both noise and speech are modeled by an autoregressive process which enables treatment of non-white and non-stationary noise scenarios. Assuming that the innovation processes are Gaussian random variables, the pdfs of speech and noise conditioned on their STP parameters are given by

$$p(\boldsymbol{w}|\boldsymbol{a}_w) \propto \mathcal{N}(\boldsymbol{w}|\boldsymbol{a}_w^T \boldsymbol{w}_q[n-1], \sigma_w^2 \boldsymbol{I})$$
$$p(\boldsymbol{x}|\boldsymbol{a}_x) \propto \mathcal{N}(\boldsymbol{x}|\boldsymbol{a}_x^T \boldsymbol{x}_p[n-1], \sigma_x^2 \boldsymbol{I})$$

where $\boldsymbol{w}_q[n-1]$ and $\boldsymbol{x}_p[n-1]$ are the Q and P past values of noise and speech, respectively. The prior model is obtained from the GMM training describes in Section 5.2. The joint pdf of the complete model is now obtained as

$$p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{a}_x, \boldsymbol{a}_w, \sigma_w^2, \sigma_x^2) = p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{a}_x, \boldsymbol{a}_w, \sigma_w^2, \sigma_x^2) p(\boldsymbol{x}|\boldsymbol{a}_x, \sigma_x^2) p(\boldsymbol{w}|\boldsymbol{a}_w, \sigma_w^2) p(\boldsymbol{a_x}) p(\boldsymbol{a}_w) p(\sigma_x^2) p(\sigma_w^2)$$

(5.11)

where $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{a}_x, \boldsymbol{a}_w, \sigma_w^2, \sigma_x^2)$ is the likelihood function, $p(\boldsymbol{a}_{\boldsymbol{x}}), p(\boldsymbol{a}_w)$ are the pre-trained GMMs and $p(\sigma_x^2), p(\sigma_w^2)$ are the *a priori* distributions of the excitation variances. The estimation is done by implementing the VB-EM algorithm as described in Chapter 2. Therefore, we define the hidden state variable as a collection of speech and noise vectors $\boldsymbol{z} = [\boldsymbol{x}\ \boldsymbol{w}]$. Furthermore we stack the parameters to a speech and noise parameter vector $\boldsymbol{\theta}_x = [\boldsymbol{a}_x\ \sigma_x^2]$ and $\boldsymbol{\theta}_w = [\boldsymbol{a}_w\ \sigma_w^2]$, respectively. Using the methods from Chapter 2, our goal is to lower bound the evidence, i.e. maximizing the free energy which is obtained as

$$\mathcal{F}(q(\boldsymbol{z}), q(\boldsymbol{\theta}_x), q(\boldsymbol{\theta}_w), \boldsymbol{y}) = \int q(\boldsymbol{z}, \boldsymbol{\theta}_x, \boldsymbol{\theta}_w) \log \left(\frac{p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}_{\boldsymbol{x}}, \boldsymbol{\theta}_w)}{q(\boldsymbol{z}, \boldsymbol{\theta}_x, \boldsymbol{\theta}_w)} \right) d\boldsymbol{z} d\boldsymbol{\theta}_x d\boldsymbol{\theta}_w \tag{5.12}$$

We now want to factorize the auxiliary distribution using the mean-field approach. Therefore, we decouple the hidden variables and the parameter distributions according to

$$q(\boldsymbol{z}, \boldsymbol{\theta}_x, \boldsymbol{\theta}_w) = q(\boldsymbol{z}) q(\boldsymbol{\theta}_x) q(\boldsymbol{\theta}_w) \tag{5.13}$$

Since noise and speech are assumed to be independent, we can follow this assumption for the auxiliary distributions. We can also assume that the LP coefficients of speech and noise as well as the excitation variances are independent, which gives the final factorized auxiliary distribution as

$$q(\boldsymbol{z}, \boldsymbol{\theta}_x, \boldsymbol{\theta}_w) = q(\boldsymbol{x}) q(\boldsymbol{w}) q(\boldsymbol{a}_x) q(\boldsymbol{a}_w) q(\sigma_x^2) q(\sigma_w^2) \tag{5.14}$$

The solutions of the optimal distributions are given by computing the expectation of the joint pdf with respect to the auxiliary distributions in an iterative way which yields the VB-EM algorithm for STP parameter estimation. Note that as a side product, the VBE-Step gives signal estimates of clean speech and noise since the VBE-Step computes the posterior distribution of the hidden states, i.e. of clean speech and noise.

5.3.2 VBE-Step

The goal in the VBE-Step is to compute the posterior distribution of the hidden state variable $\boldsymbol{z}$ which can be obtained by maximizing the free energy with respect to the auxiliary parameter distributions.

$$\log \hat{q}(\boldsymbol{z}) \propto \mathbb{E}_{q(\boldsymbol{\theta}_x, \boldsymbol{\theta}_w)} \Big\{ \log p(\boldsymbol{z}, \boldsymbol{\theta}_x, \boldsymbol{\theta}_w) \Big\} \tag{5.15}$$

By doing so, it turns out that the VBE-Step equals the Kalman filter, which was equally obtained in the method described in Chapter 4. The main difference now relies on the VBM-Step which does not compute maximum-likelihood solutions but rather implies prior distributions and maximizes the free energy of the joint pdf with previous obtained hidden state estimtes. Thus, the state estimate $\boldsymbol{z}[n|N]$ and the associated error covariance matrix $\boldsymbol{P}[n|N]$ based on the observation frame $y[n]$ where $n = 0, 1, \ldots, N-1$ are obtained in the VBE-Step. Since the Kalman filter equations were already described in Chapter 4, these equations are skipped in this chapter for better readbility.

5.3.3 VBM-Step

The heart of the new method relies in the VBM-Step which yields posterior distributions for the STP parameters of speech and noise. The information gained in the VBE-Step is extracted from the error covariance matrix of the Kalman recursions and implied in the VBM-Step. The new computed distributions allow to decrease the *variational lower bound* which leads to fast

convergence of the algorithm. The VBM-Step maximizes the free energy with given hidden state posterior from the VBE-Step. The speech and noise STP estimation can be decoupled, since independence was assumed due to the mean-field approach. This allows an elegant mathematical treatment and tractability is guaranteed.

$$\log \hat{q}(\boldsymbol{\theta}_x) \propto \mathbb{E}_{q(\boldsymbol{\theta}_w),q(\boldsymbol{z})}\Big\{\log p(\boldsymbol{z},\boldsymbol{\theta}_x,\boldsymbol{\theta}_w)p(\boldsymbol{\theta}_x)\Big\} \tag{5.16}$$

$$\log \hat{q}(\boldsymbol{\theta}_w) \propto \mathbb{E}_{q(\boldsymbol{\theta}_x),q(\boldsymbol{z})}\Big\{\log p(\boldsymbol{z},\boldsymbol{\theta}_x,\boldsymbol{\theta}_w)p(\boldsymbol{\theta}_w)\Big\} \tag{5.17}$$

A detailed treatment of the STP parameter solutions is given in the following sections. For a detailed derivation, the reader is refered to Appendix B.

Speech LP coefficients

Since the LP coefficients of speech are trained using a GMM comprising N_x mixture components, direct evaluation of the free energy is decoupled into N_x optimization problems. For every mixture component k of the prior model $p(\boldsymbol{a}_x)$, the free energy is evaluated. The final approximate distribution is then a GMM with posterior mixing coefficients and posteriori means and covariance matrix. To obtain the final estimate of the LP coefficients an MMSE estimate is used. Evaluating Eq.(5.16) for every mixture component we obtain

$$\log q(\boldsymbol{a}_x) \propto \mathbb{E}_{q(\boldsymbol{x}),q(\sigma_x^2)}\Big\{p(\boldsymbol{x}|\boldsymbol{a}_x,\sigma_x^2\boldsymbol{I})p(\boldsymbol{a}_x|k)\Big\} \tag{5.18}$$

where $p(\boldsymbol{a}_x|k)$ is the prior of $\boldsymbol{a}_x$ given component k which turns out to be a Gaussian pdf $p(\boldsymbol{a}_x|k) = \mathcal{N}(\boldsymbol{a}_x|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$. Note that Eq.(5.18) equals a product of two Gaussian densities which can easily be solved by completing the square. Thus, the result yields

$$\hat{q}(\boldsymbol{a}_x) \sim \mathcal{N}(\boldsymbol{a}_x|\hat{\boldsymbol{\mu}}_{k,\boldsymbol{a}_x},\hat{\boldsymbol{\Sigma}}_{k,\boldsymbol{a}_x}) \tag{5.19}$$

with mean and covariance matrix obtained as

$$\widehat{\boldsymbol{\Sigma}}_{k,\boldsymbol{a}_x} = \left(\mathbb{E}_{q(\sigma_x^2)}\Big\{\frac{1}{\sigma_x^2}\Big\}\sum_{n=0}^{N-1}\widehat{\boldsymbol{x}_p[n-1]\boldsymbol{x}_p[n-1]^T} + \boldsymbol{\Sigma}_k^{-1}\right)^{-1} \tag{5.20}$$

$$\widehat{\boldsymbol{\mu}}_{k,\boldsymbol{a}_x} = \widehat{\boldsymbol{\Sigma}}_{k,\boldsymbol{a}_x}\left(\mathbb{E}_{q(\sigma_x^2)}\Big\{\frac{1}{\sigma_x^2}\Big\}\sum_{n=0}^{N-1}\widehat{x[n]\boldsymbol{x}_p[n-1]}\right) \tag{5.21}$$

The expectations in Eq.(5.20) and Eq.(5.21) are easier to compute if one estimates the excitation precision which is defined as the inverese of the variance.

Responsibilities of Speech

The correct soft-assignment is only achieved if the mixing coefficients are updated accordingly. Therefore, we compute the a posteriori responsibilites and obtain the a posteriori mixing coefficients by normalization. Figure 5.4 shows an example of posterior mixing weights of a female speaker. The prior weigths are obtained after EM training and are updated via the VB-EM algorithm. The conventional CB approach would assign one cluster as the output cell whereas the GMM also incorporates neighbouring clusters which contribute to the total estimate.
The a posteriori distribution of the responsibilities is given by $q(\pi_{x,k})$ which simply takes a value

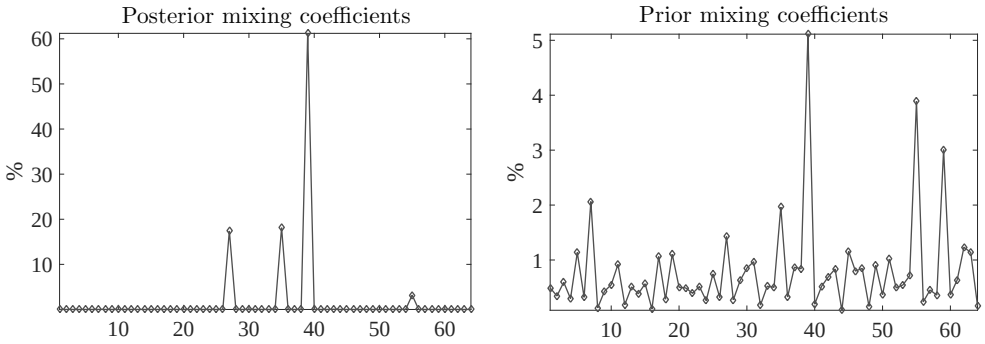

Figure 5.4: Example of computed posterior mixing coefficients from a voiced frame of a female utterance of TIMIT database performing "Did dad do academic bidding?".

in the range $\mathcal{I} = [0\ 1]$. Solving the free energy for the posteriori responsibilities yields

$$\log\ \gamma_k \propto -\frac{1}{2}\text{logdet}\ \boldsymbol{\Sigma}_k - \frac{1}{2}\Bigg(\widehat{\boldsymbol{\mu}}_{k,\boldsymbol{a}_x}^T \boldsymbol{\Sigma}_k^{-1} \widehat{\boldsymbol{\mu}}_{k,\boldsymbol{a}_x} - 2\widehat{\boldsymbol{\mu}}_{k,\boldsymbol{a}_x} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k - \text{tr}\left\{ \widehat{\boldsymbol{\Sigma}}_{k,\boldsymbol{a}_x} \boldsymbol{\Sigma}_k^{-1} \right\} \tag{5.22}$$

$$+ \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \Bigg) + \log\ \pi_k \tag{5.23}$$

The posterior mixing coefficient is obtained by normalization, i.e.

$$\rho_k = \frac{\pi_k \gamma_k}{\sum_{i=1}^{N_x} \pi_i \gamma_i} \tag{5.24}$$

The posterior distribution of the LP coefficients is then given by a GMM

$$q(\boldsymbol{a}_x) = \sum_{i=1}^{N_x} \rho_i \mathcal{N}(\boldsymbol{a}_x | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{5.25}$$

and the parameters are then given by computing the MMSE estimate of the GMM

$$\hat{\boldsymbol{a}}_x = \mathbb{E}\Big\{ \sum_{i=1}^{N_x} \rho_i \mathcal{N}(\boldsymbol{a}_x | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \Big\} = \sum_{i=1}^{N_x} \rho_i \boldsymbol{\mu}_i \tag{5.26}$$

Speech Excitation Variance

To obtain the speech excitation variances, it is mathematically easier to work with the excitation precision λ_x which is given as the inverse of the variance

$$\lambda_x = \frac{1}{\sigma_x^2} \tag{5.27}$$

The prior distribution of the precision is modeled as a Gamma distribution. This is a valid assumption since the excitation variances are non-negative and the Gamma distribution has non-negative support space. The prior is then obtained as

$$p(\lambda_x) \propto \text{Gam}\ (\lambda_x | \alpha_x, \beta_x) \tag{5.28}$$

where α_x and β_x are hyperparameters which have to be set by the experiments. Setting them to zero leads to an uniform prior which is equal to a maximum likelihood estimator. The influence of these parameters are discussed in Chapter 6. The optimal variational distributions are then

obtained by solving Eq.(5.16)

$$\log\, q(\lambda_x) \propto \mathbb{E}_{q(\mathbf{a}_x),q(\mathbf{x})}\Big\{p(\mathbf{x}|\mathbf{a}_x,\lambda_x^{-1}\mathbf{I})p(\lambda_x)\Big\} \tag{5.29}$$

From Eq.(5.29) the final result is obtained as a Gamma distribution with parameters

$$\hat{\alpha}_x = \frac{N}{2} + \alpha_x \tag{5.30}$$

$$\hat{\beta}_x = \beta_x + \frac{1}{2}\left(\sum_{n=0}^{N-1}\widehat{x}^2[n] + 2\hat{\mathbf{a}}_x\sum_{n=0}^{N-1}\widehat{x[n]\mathbf{x}_p[n-1]} + 2\hat{\mathbf{a}}_x^T\sum_{n=0}^{N-1}\widehat{\mathbf{x}_p[n-]^T\mathbf{x}_p[n-1]}\;\hat{\mathbf{a}}_x\right) \tag{5.31}$$

The excitation variance is obtained as

$$\hat{\sigma_x^2} = \frac{1}{\mathbb{E}(\lambda_x)} = \frac{\hat{\beta}_x}{\hat{\alpha}_x} \tag{5.32}$$

Noise LP coefficients

Since the LP coefficients of noise are trained using a GMM comprising N_x mixture components, direct evaluation of the free energy is decoupled into N_w maximizations. For every mixture component k of the prior model $p(\boldsymbol{a}_w)$, the free energy is evaluated. The final approximate distribution is then a GMM with posterior mixing coefficients and posteriori means and covariance matrix. To obtain the final estimate of the LP coefficients an MMSE estimate is used. Evaluating Eq.(5.17) for every mixture component we obtain

$$\log\, q(\boldsymbol{a}_w) \propto \mathbb{E}_{q(\boldsymbol{w}),q(\sigma_w^2)}\Big\{p(\boldsymbol{w}|\boldsymbol{a}_w,\sigma_w^2\boldsymbol{I})p(\boldsymbol{a}_w|k)\Big\} \tag{5.33}$$

where $p(\boldsymbol{a}_w|k)$ is the prior of $\boldsymbol{a}_w$ given component k which turns out to be a Gaussian pdf $p(\boldsymbol{a}_w|k) = \mathcal{N}(\boldsymbol{a}_w|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$. Note that Eq.(5.33) equals a product of two Gaussian densities which can easily be solved by completing the square. Thus, the result yields

$$\hat{q}(\boldsymbol{a}_w) \sim \mathcal{N}(\boldsymbol{a}_w|\hat{\boldsymbol{\mu}}_{k,\boldsymbol{a}_w},\hat{\boldsymbol{\Sigma}}_{k,\boldsymbol{a}_w}) \tag{5.34}$$

with mean and covariance matrix obtained as

$$\widehat{\boldsymbol{\Sigma}}_{k,\boldsymbol{a}_w} = \left(\mathbb{E}_{q(\sigma_w^2)}\Big\{\frac{1}{\sigma_w^2}\Big\}\sum_{n=0}^{N-1}\widehat{\boldsymbol{w}_q[n-1]\boldsymbol{w}_q[n-1]^T} + \boldsymbol{\Sigma}_k^{-1}\right)^{-1} \tag{5.35}$$

$$\widehat{\boldsymbol{\mu}}_{k,\boldsymbol{a}_w} = \widehat{\boldsymbol{\Sigma}}_{k,\boldsymbol{a}_w}\left(\mathbb{E}_{q(\sigma_w^2)}\Big\{\frac{1}{\sigma_w^2}\Big\}\sum_{n=0}^{N-1}\widehat{\boldsymbol{w}[n]\boldsymbol{w}_q[n-1]}\right) \tag{5.36}$$

The expectation in Eq.(5.35) and Eq.(5.36) are easier to compute if one estimates the excitation precision which is given as the inverese of the variance.

Responsibilities of Noise

Computing the noise responsibilities involves the same steps as for the speech process. Therefore, by maximizing the free energy and applying the MSE method one obtains the MMSE of noise

LP coefficients as

$$q(\boldsymbol{a}_w) = \sum_{i=1}^{N_w} \rho_i \mathcal{N}(\boldsymbol{a}_w | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{5.37}$$

Noise Excitation Variance

The noise excitation variances are obtained by the same way as for speech which gives

$$\hat{\alpha}_w = \frac{N}{2} + \alpha_w \tag{5.38}$$

$$\hat{\beta}_w = \beta_w + \frac{1}{2}\left(\sum_{n=0}^{N-1} \widehat{w}^2[n] + 2\hat{\mathbf{a}}_w \sum_{n=0}^{N-1} \widehat{w[n]\mathbf{w}_p[n-1]} + 2\hat{\mathbf{a}}_w^T \sum_{n=0}^{N-1} \widehat{\mathbf{w}_p[n-]^T \mathbf{w}_p[n-1]}\ \hat{\mathbf{a}}_w \right) \tag{5.39}$$

The excitation variance is obtained as

$$\hat{\sigma_w^2} = \frac{1}{\mathbb{E}(\lambda_w)} = \frac{\hat{\beta}_w}{\hat{\alpha}_w} \tag{5.40}$$

5.4 Algorithm Summary

A summary of the proposed algorithm is given below.

for *every frame* **do**
  for $i = 1 : M$ **do**
    VBE-Step
    run Kalman filter to obtain state estimates
    VBM-Step
    Speech STP: $\hat{\boldsymbol{a}}_x = \mathbb{E}\Big\{ \sum_{i=1}^{N_x} \rho_i \mathcal{N}(\boldsymbol{a}_x | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \Big\} = \sum_{i=1}^{N_x} \rho_i \boldsymbol{\mu}_i$
    $\hat{\sigma_x^2} = \frac{1}{\mathbb{E}(\lambda_x)} = \frac{\hat{\beta}_x}{\hat{\alpha}_x}$
    Noise STP: $\hat{\boldsymbol{a}}_w = \mathbb{E}\Big\{ \sum_{i=1}^{N_w} \rho_i \mathcal{N}(\boldsymbol{a}_w | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \Big\} = \sum_{i=1}^{N_w} \rho_i \boldsymbol{\mu}_i$
    $\hat{\sigma_w^2} = \frac{1}{\mathbb{E}(\lambda_w)} = \frac{\hat{\beta}_w}{\hat{\alpha}_w}$
  end
end

Algorithm 2: VB-EM algorithm for STP coefficients estimation.

Note that the variational lower bound gets decreased after every iteration, leading to guaranteed convergence. Nevertheless, a global optimum may not be achieved and the algorithm only provides local optima solutions. A block diagram of the proposed method is given below.

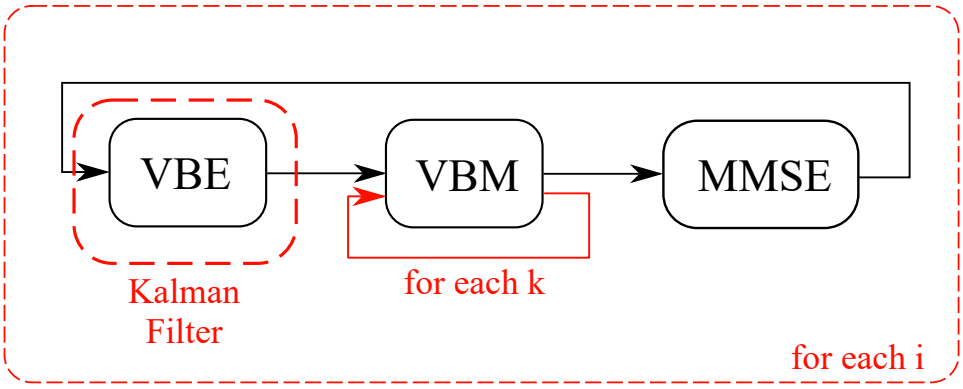

Figure 5.5: Schematic block diagram of the proposed method.

5.5 Discussion

It is very interesting to discuss the properties of the new method and compare it to conventional methods. By assuming uniform priors one obtains the well-known ML solution which is similar to the Yule-Walker equations. The difference is given by the fact that the speech and noise statistics are extracted from the Kalman filter posterior densities. If a prior distribution is given, the solutions are obtained as *modified* Yule-Walker equations. The prior covariance matrix matrix acts on the result as an additive term. Assuming a huge covariance matrix which is equal to a large uncertainty in the prior model, the estimator relies more on the observed data. If there is a good prior available, the estimator relies more on the prior and less on the data.
For example, in voiced frames at low SNR the STP parameters are heavily corrupted and the Kalman filter would give bad results. Therefore, a good prior in voiced frames can extract more robust estimates and allows better performance. This low SNR scenario was already adressed in the ML approach where the Kalman filter was not able to improve the performance. The initialization was obtained by computing third-order cumulants of speech and assuming an additive white noise model, which is not valid in real scenarios. The new approach can reduce the bias due to the hyperparameters of the Gamma distributions since the Kalman filter is more sensitive to the excitation variances. The reason is given by the fact that the excitation variances model the process noise of the Kalman filter, i.e. the process covariance matrix. Biased estimates lead to poor behaviour of the Kalman filter due to the wrong modelling of the process noise covariance matrix. Compared to the CB method, the computational complexity could be reduced since no full-search algorithm needs to be implemented. Also, there is no need to constrain the exctiation variances since the Gamma distribution has positive support and negative excitation variances cannot occur. Table 5.2 gives an overview of the proposed method compared to the ML solution. Note, that without a pre-trained GMM, the porposed method equals the ML method. Thus, the proposed method could be understand as a generalization of the Kalman-filter based framework.

Table 5.2: Comparison of ML-KF approach and proposed method.

ML method	**Proposed method**
LP coefficients:	LP coefficients
$\hat{\mathbf{a}}_x = -\left(\sum_{n=0}^{N-1} \mathbf{x}_p[n-1]\mathbf{x}_p^T[n-1]\right)^{-1} \sum_{n=0}^{N-1} \mathbf{x}_p[n-1]x[n]$	$\widehat{\boldsymbol{\mu}}_{k,\boldsymbol{a}_x} = \widehat{\boldsymbol{\Sigma}}_{k,\boldsymbol{a}_x}\left(\mathbb{E}_{q(\sigma_x^2)}\left\{\frac{1}{\sigma_x^2}\right\}\sum_{n=0}^{N-1} x[n]\widehat{\boldsymbol{x}_p[n}-1]\right)$
	$\widehat{\boldsymbol{\Sigma}}_{k,\boldsymbol{a}_x} = \left(\mathbb{E}_{q(\sigma_x^2)}\left\{\frac{1}{\sigma_x^2}\right\}\sum_{n=0}^{N-1} \boldsymbol{x}_p[n-\widehat{1]\boldsymbol{x}_p}[n-1]^T + \boldsymbol{\Sigma}_k^{-1}\right)^{-1}$
Excitation variance:	Excitation variance:
$\sigma_x^2 = \frac{1}{N}\sum_{n=0}^{N-1}\left(x^2[n] + \hat{\mathbf{a}}_x^T\mathbf{x}_p[n-1]x[n]\right)$	$\hat{\alpha}_x = \frac{N}{2} + \alpha_x$
	$\hat{\beta}_x = \beta_x + \frac{1}{2}\left(\sum_{n=0}^{N-1}\hat{x}^2[n] + 2\hat{\mathbf{a}}_x\sum_{n=0}^{N-1} x[n]\widehat{\mathbf{x}_p[n}-1]\right.$
	$\left.+2\hat{\mathbf{a}}_x^T\sum_{n=0}^{N-1}\mathbf{x}_p[n-]\widehat{^T\mathbf{x}_p}[n-1]\hat{\mathbf{a}}_x\right)$
	$\sigma_x^2 = \frac{\hat{\beta}_x}{\hat{\alpha}_x}$

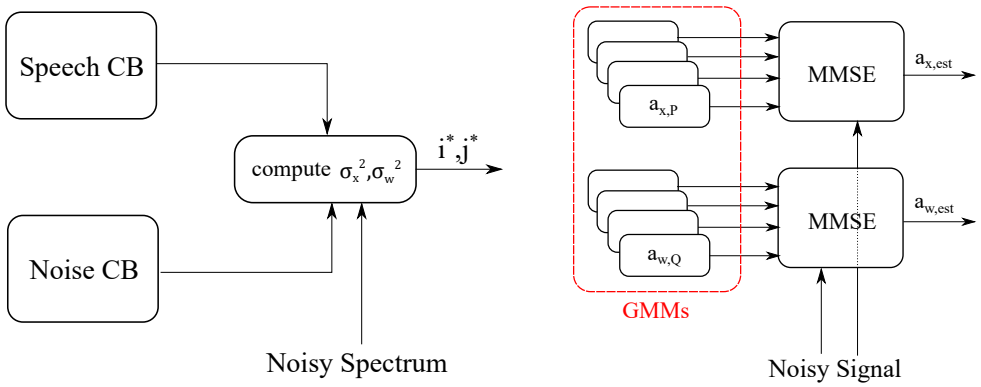

Figure 5.6: Comparison of the CB-based enhancement (left) and the proposed enhancement scheme (right)-

Figure 5.6 shows two block diagrams of the enhancement process of the CB-based approach and the proposed method. Note that the CB method goes through the whole codebook by minimizing the Itakuro-Saito distance for each possible pair. If the excitation variances were found, the optimal codebook entries are chosen by minimizing the likelihood function. The new method does not rely on a full-search procedure but computes MMSE estimates using all GMM mixtures at once. Furthermore, the excitation variances are computed itertively by using the estimates of clean speech, noise and LP coefficients from the last VB-EM iteration.

6 Results

This chapter discusses the main results of this thesis. Therefore, the proposed method was investigated for different noise scenarios and compared to reference methods. The chapter is organized as follows; Section 6.1. describes the setup of the experiments and used evaluation criteria. Section 6.2. presents the obtained results of the proposed methods and reference methods.

6.1 Experimental Setup

6.1.1 Evaluation Metrics

Speech Quality

To evaluate the speech quality of the enhanced signal, two metrics are used. First, the *Perceptual Evaluation of Speech Quality* (PESQ) which is recommended by ITU-T [45], which is based on a cognitive model of quality. The PESQ metric range is bounded within the interval $[-0.5\ 4.5]$ where 4.5 corresponds to the maximum achivable quality and -0.5 corresponds to the minimum, respectively. Most modern speech enhancement systems use the PESQ score as a quality metric although it is very sensitive to over-harmonization of the signal. This can lead to better PESQ results using phase estimation methods than obtained by the clean phase.
The second quality metric is the *segmental signal-to-noise ratio* (SSNR) which is one of the most-known and most used objective scores in speech processing. The SSNR calculation is done for short-time frames and the total output is obtained by averaging the frame SNRs.

$$\mathrm{SSNR} = \frac{1}{M}\sum_{m=0}^{M-1}\log\left(\sum_{n=Nm}^{n=Nm+N-1}\left(\frac{\sum_{n=0}^{N-1}x^2[n]}{\sum_{n=0}^{N-1}(x[n]-y[n])^2}\right)\right) \tag{6.1}$$

where M is the number of frames and N is the framelength, respectively.

Speech Intelligibility

To predict speech intelligibility, the Short Time Objective Intelligibility metric was proposed in [46]. Many speech enhancement methods degrade speech intelligibility and in earlier literature only speech quality scores were reported. STOI is computed by segmenting input speech in 15 one-third bands where the lowest center frequency is defined by 150 Hz. Silent frame are excluded for the evaluation and normalizing and temporal smoothing is applied to the voiced frames. The final STOI measure is obtained by averaging a short-time sample correlation coefficient along time and sub-bands. The STOI score ranges from 0 (worse) to 1 (best).

6.1.2 Databases

Speech Database

For all experiments speech signals from the TIMIT database [47] were taken to verify the proposed method. The database was designed by a joint effort among Massachusetts Institute of Technology (MIT), SRI International and Texas Instruments, Inc. (TI). The corpus contains broadband speech form 630 speakers of eight different dialects of American English containing a total of 6300 speech files sampled at a sampling frequency of 16kHz. For the experiments, 50 utterances of the corpus were randomely chosen (25 male and 25 female).

Noise Database

For the experiments the noise recordings were taken from the NOISEX-92 database [48]. Stationary and non-stationary noise types were chosen to prove the capabilities of the proposed method. For non-stationary types babble noise and factory noise were chosen. Babble noise is a recording of people talking and simulates scenarios like restaurants or cocktail-party problems. Factory noise simulates background noise of production sides and factories. Furthermore white and pink noise were used. White noise is one of the most important noise types since it models many phenomena in engineering, like measurement noise or the random movement of electrons. The main property of white noise is its flat power spectrum. Pink noise is a low-frequency noise type which can be obtained as a low-pass filtered version of white noise.

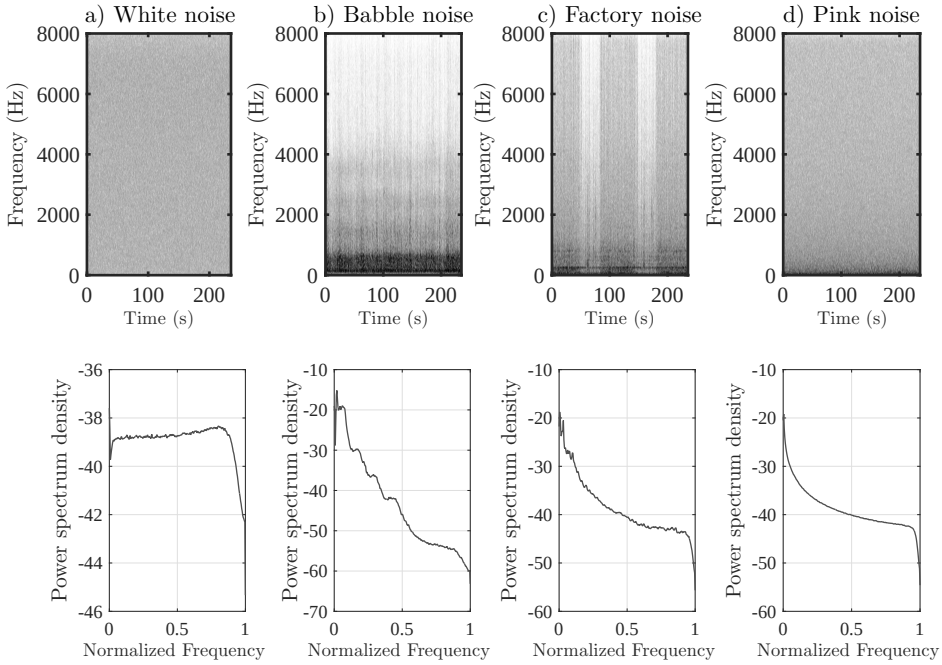

Figure 6.1: Spectrograms (top) and PSDs obtained from Welchs method [49] (bottom) of different noise types.

6.2 Results

This section highlights the main results of the proposed method. The overall estimation results for different noise types are shown as well as the behavior of the algorithm for different design parameters, such as the number of GMM components and the initialization of the hyperparameters of the excitation prior.

6.2.1 Influence of Mixture Components

The total number of mixture components of the prior GMMs is an important parameter which has to be chosen carefully. Many mixture components increase the computational complexity, not only at the training stage, but also during the enhancement process. On the other hand, too few components may lead to worse output speech and therefore it is questionable how to set the number of components. Figure 6.2. shows PESQ and STOI improvement for different speech mixtures ranging in the interval $\mathcal{I} = [8\ 32\ 64\ 128\ 512]$. The PESQ and STOI values quickly saturates for $N_x \geq 32$ which gives the optimum number of coefficients for superior performance. For all further experiments, N_x was set to the optimal value obtained from these experiments. Since no increase in PESQ and STOI was obtained for higher mixtures, the computational complexity could be drastically reduced. In the CB approach it was reported that a 10-bit codebook was trained for the enhancement which equals to $N_x = 1024$ components [8]. This leads to an enormous complexity since a full search through all codebook combinations has to be undertaken. The GMM approach is able to reduce this complexity issue which results in much faster overall performance.

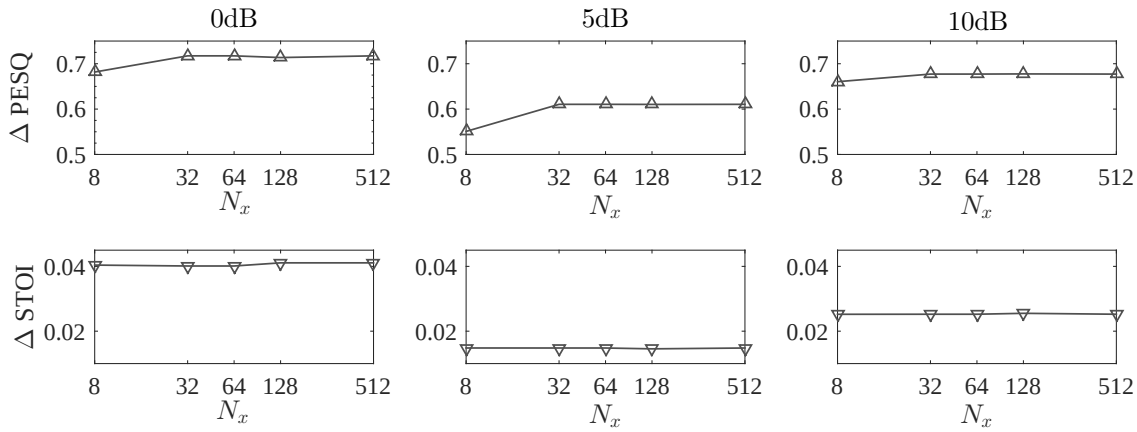

Figure 6.2: Influence of mixture components averaged over 25 utterances. PESQ improvement (top) for different signal-to-noise ratios and STOI improvement (bottom).

6.2.2 Influence of Excitation Hyperparameters

Since the speech and noise excitation variance priors are modeled by a Gamma distribution, it is doubtful how to chose the hyperparameters. Therefore, simulations were carried out to obtain the hyperparameter influence with respect to the enhanced speech by computing PESQ and STOI values for different setups. The standard maximum-likelihood solution is obtained by setting the hyperparameters to zero. Experiments showed that this leads to an overestimation of the variances and therefore speech is degraded. To circumvent this issue, hyperparameters can reduce the bias and therefore contribute to better speech quality and intelligibility. Nevertheless caution must be taken, since wrongly chosen values may lead to underestimation of the excitation variances. Table 6.1. shows different hyperparameter settings. Experiments were conducted and averaged over 25 utterances of the TIMIT database.

Table 6.1: Experimental setup for hyperparameter investigation.

	Setup	
#1	$b_x = 0$	$a_x \in [10,\ 10^6]$
#2	$b_w = 0$	$a_w \in [10,\ 10^6]$

Figure 6.3 depicts results for setups #1 and #2 and show that an increase of a_x results in better speech quality and intelligibility but drops if the value is chosen too big which reflects the underestimation of the excitation variance. By increasing a_w the PESQ value decreases and was set to zero for further experiments. The STOI value saturates for larger values. Since the VB-EM noise parameter initialization is obtained from the minimum statistics approach a good initialization is provided. The speech initialization is obtained from the observed noisy speech frame which is worse for low SNR regions, especially in voiced regions. This yields a biased initialization of excitation variances which can be compensated by setting appropriate hyperparameters of the Gamma distribution.

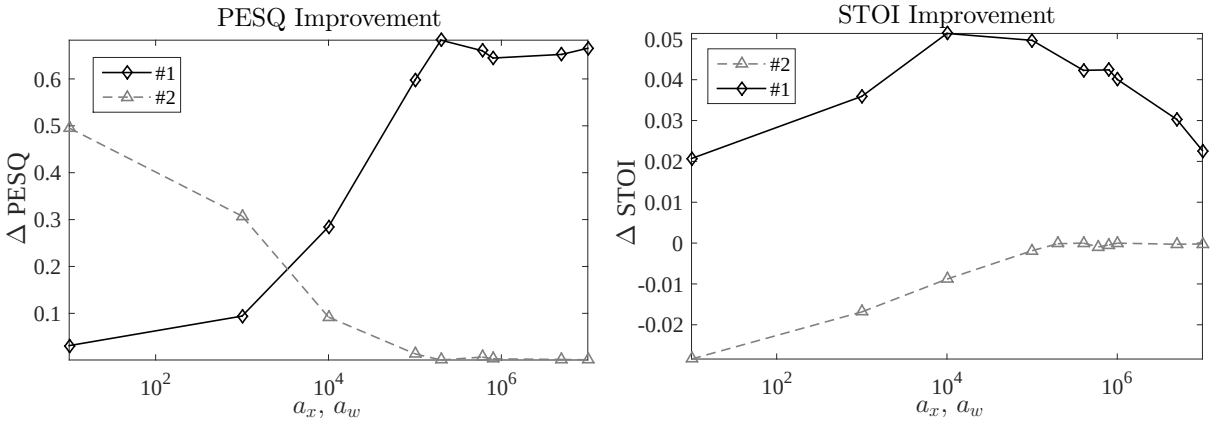

Figure 6.3: PESQ improvement (left) and STOI improvement (right) for setup #1 (black solid line) and setup #2 (dashed gray line) averaged over 25 utterances from the TIMIT database for a 0dB white noise scenario.

6.2.3 Evaluation

The evaluation was done by using the optimal GMM size and the optimal hyperparameter values obtained from the previous section which leads the best results. The performance was measured by computing PESQ, STOI and SSNR scores over 50 utterances of the TIMIT database for different noise scenarios at different SNR levels. As an upper bound, a ground truth Kalman filter was implemented which has knowledge of the clean speech and noise STP parameters. The method was compared to the well-known LSA method [50], the ML-based Kalman filter method (ML-KF) [6], proposed LSA-based method using a GMM Kalman filter postprocessor (GMM-KF + LSA) and a speaker-dependent speech enhancement (SDSE) method proposed in [51]. The speech LP order was set to $P = 10$ for all experiments and the noise LP order was set

Table 6.2: Overview of the setup for evaluation.

Evaluation Setup	
SNR	$0 - 5 - 10$ [dB]
noisetypes	white, babble, pink, factory [48]
# utterances	50
sampling frequency f_s	16kHz
framelength N	512 samples
window type	Hamming
frame shift	256 samples

according to Table 6.3.

Table 6.3: Model orders for different noise scenarios.

Noise type	Settings
white	$Q = 4$
babble	$Q = 10$
factory	$Q = 8$
pink	$Q = 6$

An overall improvement and outperformance at low SNR of PESQ and SSNR was obtained for the new method. In non-stationary enviroments like babble noise, speech quality was improved but speech intelligbility was reduced. The reason for this behaviour may be the initialization of the Kalman filter which is crucial for good overall estimation performance. Since the initialization of noise was done using the minimum statistics approach, a mismatch occurs in non-stationary enviroments which are not robust estimated by the minimum statistics approach. On the other hand, the initialization of the speech state space model was obtained from the noisy speech samples. This leads to a mismatch at low SNR regions were voiced frames may be heavily masked by noise. A bigger problem are unvoiced regions which cannot be precisley tracked by the Kalman filter. The figures on the next pages show delta values of the evaluation metrics for different noise scenarios at different SNRs.

6.2.4 White Noise

For white noise, the proposed method outperforms both the LSA and the ML-KF method in terms of PESQ and SSNR improvement which yields superior speech quality. For low SNR scenarios the proposed method also outperformed LSA and ML-KF in terms of STOI.

PESQ Improvement

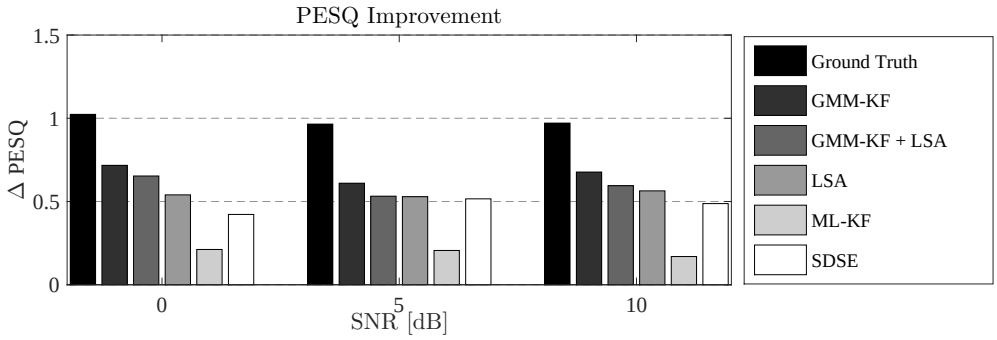

Figure 6.4: PESQ improvement averaged over 50 utterances for white noise at three different SNRs.

STOI Improvement

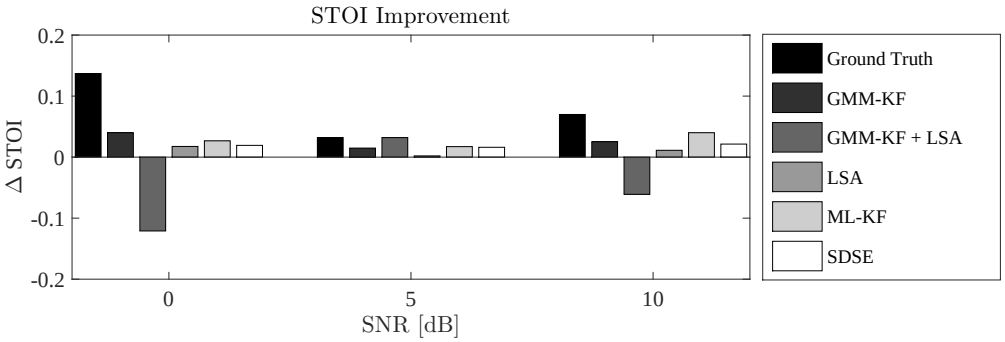

Figure 6.5: STOI improvement averaged over 50 utterances for white noise at three different SNRs.

SSNR Improvement

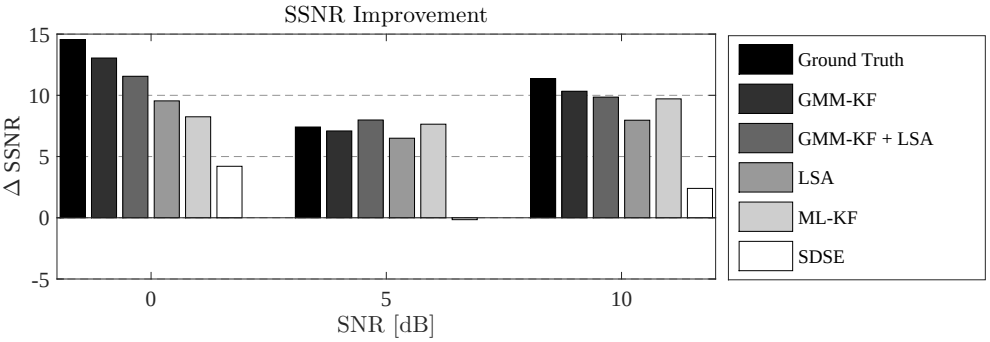

Figure 6.6: SSNR improvement averaged over 50 utterances for white noise at three different SNRs.

6.2.5 Babble Noise

For the babble noise scenario the proposed method was able to give fairly good PESQ and SSNR improvement in high-SNR regions. It is interesting to note that for the 0dB case the proposed method yields best results. Also, by implementing the GMM-KF method as a postprocessing unit an increase in PESQ and SSNR was obtained. In terms of speech intelligibility the proposed method was not able to further improve the STOI score.

PESQ Improvement

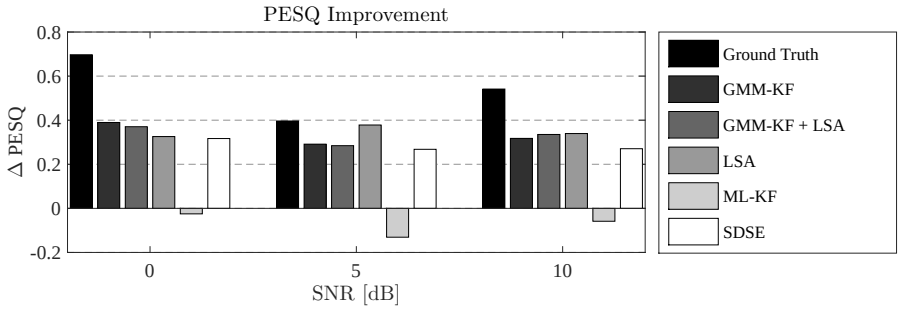

Figure 6.7: PESQ improvement averaged over 50 utterances for babble noise at three different SNRs.

STOI Improvement

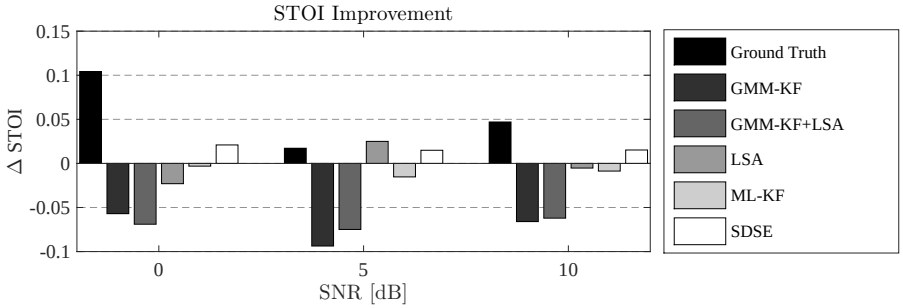

Figure 6.8: STOI improvement averaged over 50 utterances for babble noise at three different SNRs.

SSNR Improvement

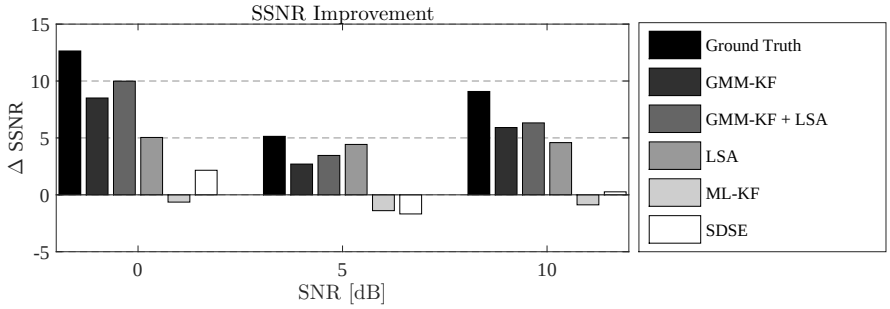

Figure 6.9: SSNR improvement averaged over 50 utterances for babble noise at three different SNRs.

6.2.6 Factory Noise

at low SNR scenarios the proposed method gives improved PESQ and SSNR, especially when used as a postprocessing unit. Nevertheless, decrease in STOI was obtained but the algorithm tends to be more robust compared to the babble noise scenario.

PESQ Improvement

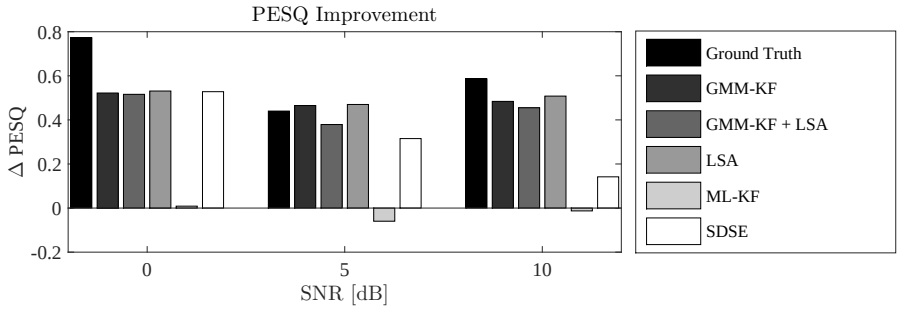

Figure 6.10: PESQ improvement averaged over 50 utterances for factory noise at three different SNRs.

STOI Improvement

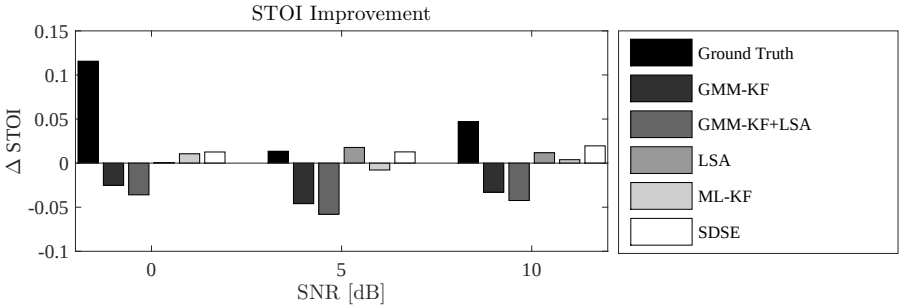

Figure 6.11: STOI improvement averaged over 50 utterances for factory noise at three different SNRs.

SSNR Improvement

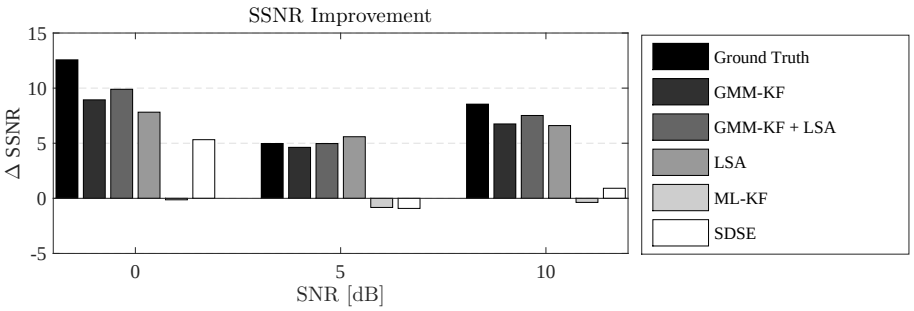

Figure 6.12: SSNR improvement averaged over 50 utterances for factory noise at three different SNRs.

6.2.7 Pink Noise

Pink noise showed superior improvement at low SNR regions for PESQ and SNR. Also STOI increased for the 0dB scenario and outperformed LSA method.

PESQ Improvement

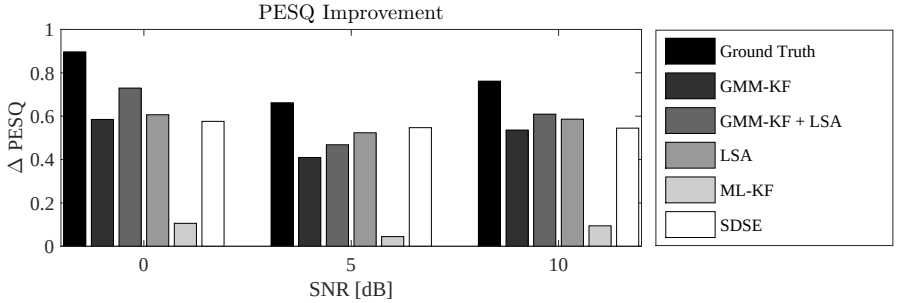

Figure 6.13: PESQ improvement averaged over 50 utterances for pink noise at three different SNRs.

STOI Improvement

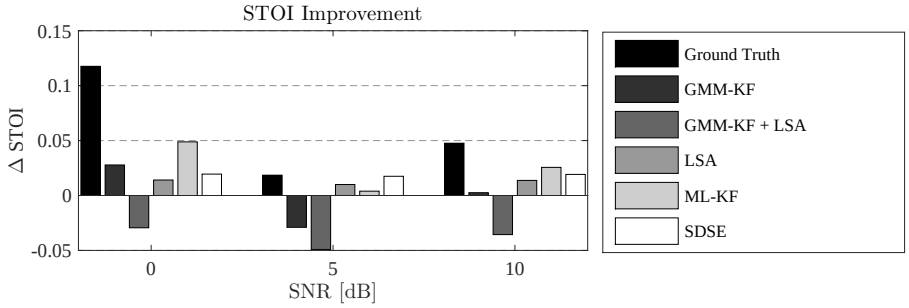

Figure 6.14: STOI improvement averaged over 50 utterances for pink noise at three different SNRs.

SSNR Improvement

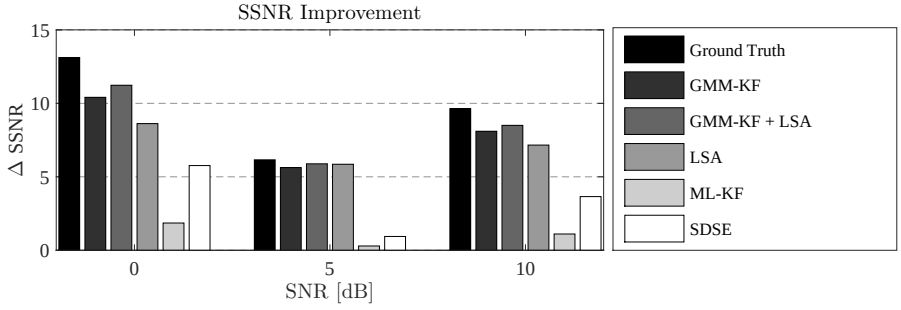

Figure 6.15: SSNR improvement averaged over 50 utterances for pink noise at three different SNRs.

6.2.8 Spectrogram Examples

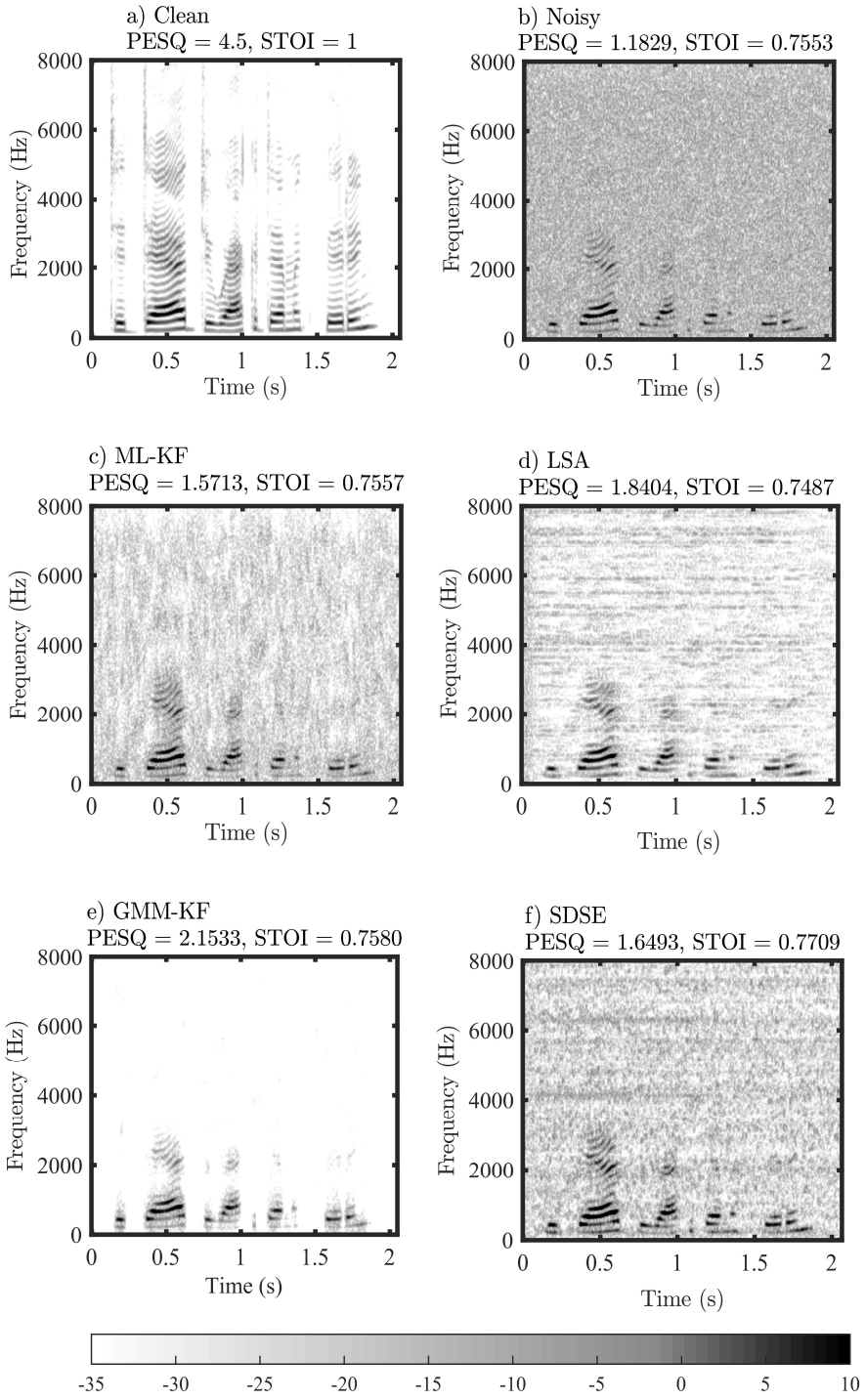

Figure 6.16: Spectograms and corresponding PESQ and STOI values of an female speaker from TIMIT database performing "Did dad do academic bidding?": a) clean speech, b) noisy speech corrupted with 0dB white noise, c) enhanced signal using ML-KF method, d) enhanced signal using the LSA method, e) enhanced signal using GMM-KF approach with $N_x = 512$ speech components and $N_w = 128$ noise components, f) enhanced signal using LSA method with GMM-KF post-processing.

7 Conclusion and Outlook

This thesis presented a new approach in kowledge-based speech enhancement which makes use of GMMs contrary to previos works which focused on codebook-based approaches. Additionally, a variational Bayesian framework was established which yields an iterative framework based on the EM algorithm approximating the posterior distributions of the complete data and the parameters of interest. The overall performance of the new approach was justified by experiments for different noise types - stationary as well as non-stationary - and different SNRs. As an upper bound served the Kalman filter with known speech and noise parameters. PESQ, SSNR and STOI improvement was observed for stationary noise types. In non-stationary scenarios PESQ and SSNR improvement was observed but for some noise types STOI could not be raised. A reduction in complexity was also achieved since the GMM could enhance speech with just a few mixture components of speech and noise which yields a fast implementation. Also, the influence of design parameters such as the hyperparameters was investigated which showed an improvement at low SNR regions since the hyperparameters can reduce the bias in the estimated excitation variance giving better results in PESQ and STOI. It should be noted that some high-frequency portions of the speech signals could not be recovered which leads to some sort of buzz in the output speech. Nevertheless, artefacts like musical noise did not occur to an high extend, especially at sationary noise scenarios. Future work may also include the Bayesian estimation of the model order using either sparse Bayesian learning methods or information-theoretic criteria. Also the mean-field approach which assumes independence between hidden states and parameters could be swapped by a conditional factorization of the auxiliary distributions.

A

Derivation of the Variational Lower Bound

Using Bayes law we can formulate the posterior distribution of the hidden variables

$$p(\boldsymbol{a}, \boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\theta})p(\boldsymbol{a}, \boldsymbol{\theta})}{p(\boldsymbol{y})} \tag{A.1}$$

We now want to derive a variational lower bound for the given model in Eq.(A.1) which leads us to a minimization of the Kullback-Leibler distance between the real posterior distribution in Eq.(A.1) and a variational distribution of simpler, tractable form. To start with the derivation, we define the Kullback-Leibler distance between two distributions as

$$\mathcal{D}_{\mathrm{KL}}(p||q) = \int p(x)\log\left(\frac{p(x)}{q(x)}\right) dx \tag{A.2}$$

Due to Gibbs inequality, the Kullback-Leibler distance always satisfies the condition $\mathcal{D}_{\mathrm{KL}} \geq 0$ and is zero if and only if $q(x) = p(x)$.
We now move on with the derivation of the variational lower bound. By inserting the posterior and variational distribution in Eq.(1.5) we obtain

$$\mathcal{D}_{\mathrm{KL}}(q||p) = \int q(\boldsymbol{a}, \boldsymbol{\theta})\log\left(\frac{q(\boldsymbol{a}, \boldsymbol{\theta})}{p(\boldsymbol{a}, \boldsymbol{\theta}|\boldsymbol{y})}\right) d\boldsymbol{a}d\boldsymbol{\theta} \tag{A.3}$$

By inserting Eq.(1.4) into Eq.(1.6) we further obtain

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}(q||p) &= \int q(\boldsymbol{a}, \boldsymbol{\theta})\log\left(\frac{q(\boldsymbol{a}, \boldsymbol{\theta})p(\boldsymbol{y})}{p(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\theta})p(\boldsymbol{a}, \boldsymbol{\theta})}\right) d\boldsymbol{a}d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{a}, \boldsymbol{\theta})\log\left(\frac{q(\boldsymbol{a}, \boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\theta})p(\boldsymbol{a}, \boldsymbol{\theta})}\right) d\boldsymbol{a}d\boldsymbol{\theta} + \log\, p(\boldsymbol{y}) \\
&\geq \int q(\boldsymbol{a}, \boldsymbol{\theta})\log\left(\frac{q(\boldsymbol{a}, \boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\theta})p(\boldsymbol{a}, \boldsymbol{\theta})}\right) d\boldsymbol{a}d\boldsymbol{\theta}
\end{aligned} \tag{A.4}$$

Therefore we can define the variational lower bound (VLB) as

$$\mathcal{L}(q) = \int q(\boldsymbol{a}, \boldsymbol{\theta})\log\left(\frac{p(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\theta})p(\boldsymbol{a}, \boldsymbol{\theta})}{q(\boldsymbol{a}, \boldsymbol{\theta})}\right) d\boldsymbol{a}d\boldsymbol{\theta} \tag{A.5}$$

B

Derivations of Variational distributions

B.1 LP coefficients of Speech Process

$$\begin{aligned}\ln\, q(\boldsymbol{a}_x|k) &= \mathbb{E}_{q(\boldsymbol{x})q(\lambda_x)}\Big\{\ln\, p(\boldsymbol{x}|\boldsymbol{a}_x,\lambda_x)p(\boldsymbol{a}_x|k)\Big\}\\ &= \mathbb{E}_{q(\boldsymbol{x})q(\lambda_x)}\Big\{C-\frac{\lambda_x}{2}\Big(\boldsymbol{x}^T\boldsymbol{x}-2\boldsymbol{a}_x^T\boldsymbol{x}_p\boldsymbol{x}^T+\boldsymbol{a}_x^T\boldsymbol{x}_p\boldsymbol{x}_p^T\boldsymbol{a}_x\Big)\\ &-\frac{1}{2}\Big(\boldsymbol{a}_x^T\boldsymbol{\Sigma}_{x,k}^{-1}\boldsymbol{a}_x-2\boldsymbol{\mu}_{x,k}^T\boldsymbol{\Sigma}_{x,k}^{-1}\boldsymbol{a}_x\Big)\Big\}\end{aligned}$$

where $C=\frac{N}{2}\log\,\lambda_x - N\log\,2\pi-\frac{1}{2}\boldsymbol{\mu}_{x,k}^T\boldsymbol{\Sigma}_{x,k}^{-1}\boldsymbol{\mu}_{x,k}$ is treated as a constant since its independent of $\boldsymbol{a}_x$. By solving the above equation with respect to the expectation operator and completing the square we obtain the final solution.

$$\hat{q}(\boldsymbol{a}_x)\sim\mathcal{N}(\boldsymbol{a}_x|\hat{\boldsymbol{\mu}}_{k,\boldsymbol{a}_x},\hat{\boldsymbol{\Sigma}}_{k,\boldsymbol{a}_x})$$

with mean and covariance matrix obtained as

$$\widehat{\boldsymbol{\Sigma}}_{k,\boldsymbol{a}_x}=\left(\mathbb{E}_{q(\lambda_x)}\Big\{\frac{\lambda_x}{2}\Big\}\sum_{n=0}^{N-1}\widehat{\boldsymbol{x}_p[n-1]\boldsymbol{x}_p[n-1]^T}+\boldsymbol{\Sigma}_k^{-1}\right)^{-1}$$

$$\widehat{\boldsymbol{\mu}}_{k,\boldsymbol{a}_x}=\widehat{\boldsymbol{\Sigma}}_{k,\boldsymbol{a}_x}\left(\mathbb{E}_{q(\lambda_x)}\Big\{\frac{\lambda_x}{2}\Big\}\sum_{n=0}^{N-1}\widehat{x[n]\boldsymbol{x}_p[n-1]}\right)$$

B.2 LP coefficients of Noise Process

$$\begin{aligned}\ln\, q(\boldsymbol{a}_w|k) &= \mathbb{E}_{q(\boldsymbol{w})q(\lambda_w)}\Big\{\ln\, p(\boldsymbol{w}|\boldsymbol{a}_w,\lambda_w)p(\boldsymbol{a}_w|k)\Big\}\\ &= \mathbb{E}_{q(\boldsymbol{w})q(\lambda_w)}\Big\{C-\frac{\lambda_w}{2}\Big(\boldsymbol{w}^T\boldsymbol{w}-2\boldsymbol{a}_w^T\boldsymbol{w}_q\boldsymbol{w}^T+\boldsymbol{a}_w^T\boldsymbol{w}_q\boldsymbol{w}_q^T\boldsymbol{a}_w\Big)\\ &-\frac{1}{2}\Big(\boldsymbol{a}_w^T\boldsymbol{\Sigma}_{w,k}^{-1}\boldsymbol{a}_w-2\boldsymbol{\mu}_{w,k}^T\boldsymbol{\Sigma}_{w,k}^{-1}\boldsymbol{a}_w\Big)\Big\}\end{aligned}$$

where $C=\frac{N}{2}\log\,\lambda_w - N\log\,2\pi-\frac{1}{2}\boldsymbol{\mu}_{w,k}^T\boldsymbol{\Sigma}_{w,k}^{-1}\boldsymbol{\mu}_{w,k}$ is treated as a constant since its independent of $\boldsymbol{a}_w$. By solving the above equation with respect to the expectation operator and completing the square we obtain the final solution.

$$\hat{q}(\boldsymbol{a}_w)\sim\mathcal{N}(\boldsymbol{a}_w|\hat{\boldsymbol{\mu}}_{k,\boldsymbol{a}_w},\hat{\boldsymbol{\Sigma}}_{k,\boldsymbol{a}_w})$$

with mean and covariance matrix obtained as

$$\widehat{\boldsymbol{\Sigma}}_{k,\boldsymbol{a}_w} = \left(\mathbb{E}_{q(\lambda_w)}\Big\{ \frac{\lambda_w}{2} \Big\} \sum_{n=0}^{N-1} \widehat{\boldsymbol{w}_q[n-1]\boldsymbol{w}_q[n-1]^T} + \boldsymbol{\Sigma}_k^{-1} \right)^{-1}$$

$$\widehat{\boldsymbol{\mu}}_{k,\boldsymbol{a}_w} = \widehat{\boldsymbol{\Sigma}}_{k,\boldsymbol{a}_w} \left(\mathbb{E}_{q(\lambda_w)}\Big\{ \frac{\lambda_w}{2} \Big\} \sum_{n=0}^{N-1} \widehat{w[n]\boldsymbol{w}_q[n-1]} \right)$$

B.3 Excitation Precision of Speech Process

$$\begin{aligned} \ln\, q(\lambda_x) &= \mathbb{E}_{q(\boldsymbol{x})q(\boldsymbol{a}_x)}\Big\{ \ln\, p(\boldsymbol{x}|\boldsymbol{a}_x, \lambda_x) p(\lambda_x) \Big\} \\ &= \mathbb{E}_{q(\boldsymbol{x})q(\boldsymbol{a}_x)}\Big\{ \frac{N}{2} \log\, \lambda_x - \frac{\lambda_x}{2}\Big(\boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{a}_x^T\boldsymbol{x}_p\boldsymbol{x}^T + \boldsymbol{a}_x^T\boldsymbol{x}_p\boldsymbol{x}_p^T\boldsymbol{a}_x \Big) + (\alpha-1)\log\, \lambda_x - \beta\lambda_x + C \Big\} \end{aligned}$$

where $C = \frac{N}{2}\log\,\alpha - \frac{N}{2}\log\,2\pi$ is treated as a constant since its independent of λ_x. By rearranging terms and using the MMSE estimate of the LP coefficients we observe that the auxiliary posterior yields a Gamma distribution given as

$$\hat{\alpha}_x = \frac{N}{2} + \alpha_x \tag{B.1}$$

$$\hat{\beta}_x = \beta_x + \frac{1}{2}\left(\sum_{n=0}^{N-1} \widehat{x}^2[n] + 2\hat{\mathbf{a}}_x \sum_{n=0}^{N-1} \widehat{x[n]\mathbf{x}_p[n-1]} + 2\hat{\mathbf{a}}_x^T \sum_{n=0}^{N-1} \widehat{\mathbf{x}_p[n-]^T\mathbf{x}_p[n-1]}\ \hat{\mathbf{a}}_x \right) \tag{B.2}$$

where the excitation variance is obtained as the inverse of the expectation of the precision

$$\hat{\sigma_x^2} = \frac{1}{\mathbb{E}(\lambda_x)} = \frac{\hat{\beta}_x}{\hat{\alpha}_x} \tag{B.3}$$

B.4 Excitation Precision of Noise Process

$$\begin{aligned} \ln\, q(\lambda_w) &= \mathbb{E}_{q(\boldsymbol{w})q(\boldsymbol{a}_w)}\Big\{ \ln\, p(\boldsymbol{w}|\boldsymbol{a}_w, \lambda_w) p(\lambda_w) \Big\} \\ &= \mathbb{E}_{q(\boldsymbol{w})q(\boldsymbol{a}_w)}\Big\{ \frac{N}{2} \log\, \lambda_w - \frac{\lambda_w}{2}\Big(\boldsymbol{w}^T\boldsymbol{w} - 2\boldsymbol{a}_w^T\boldsymbol{w}_q\boldsymbol{w}^T + \boldsymbol{a}_w^T\boldsymbol{w}_q\boldsymbol{w}_q^T\boldsymbol{a}_w \Big) + (\alpha-1)\log\, \lambda_w - \beta\lambda_w + \end{aligned}$$

where $C = \frac{N}{2}\log\,\alpha - \frac{N}{2}\log\,2\pi$ is treated as a constant since its independent of λ_w. By rearranging terms and using the MMSE estimate of the LP coefficients we observe that the auxiliary posterior yields a Gamma distribution given as

$$\hat{\alpha}_w = \frac{N}{2} + \alpha_w \tag{B.4}$$

$$\hat{\beta}_w = \beta_w + \frac{1}{2}\left(\sum_{n=0}^{N-1} \widehat{w}^2[n] + 2\hat{\mathbf{a}}_w \sum_{n=0}^{N-1} \widehat{w[n]\mathbf{w}_q[n-1]} + 2\hat{\mathbf{a}}_w^T \sum_{n=0}^{N-1} \widehat{\mathbf{w}_q[n-1]^T\mathbf{w}_q[n-1]}\hat{\mathbf{a}}_w \right) \tag{B.5}$$

where the excitation variance is obtained as the inverse of the expectation of the precision

$$\hat{\sigma_w^2} = \frac{1}{\mathbb{E}(\lambda_w)} = \frac{\hat{\beta}_w}{\hat{\alpha}_w} \tag{B.6}$$

Bibliography

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, Inc., 2nd ed., 2013.

[2] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, pp. 197–210, Jun 1978.

[3] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Transactions on Signal Processing*, vol. 39, pp. 795–805, Apr 1991.

[4] A. M. E. Masgrau, J. Salavedra and A. Ardanuy, "Speech enhancement by adaptive wiener filtering based on cumulant AR modeling," *Speech Processing in Adverse Conditions*, pp. 143–146, 1992.

[5] E. Weinstein, A. V. Oppenheim, M. Feder, and J. R. Buck, "Iterative and sequential algorithms for multisensor signal enhancement," *IEEE Transactions on Signal Processing*, vol. 42, pp. 846–859, Apr 1994.

[6] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential kalman filter-based speech enhancement algorithms," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 373–385, Jul 1998.

[7] T. V. Sreenivas and P. Kirnapure, "Codebook constrained wiener filtering for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 383–389, Sep 1996.

[8] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 163–176, Jan 2006.

[9] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 441–452, Feb 2007.

[10] T. Rosenkranz and H. Puder, "Improving robustness of codebook-based noise estimation approaches with delta codebooks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1177–1188, May 2012.

[11] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[13] C. Andrieu, N. D. Freitas, and et al., "An introduction to mcmc for machine learning," 2003.

[14] C. M. Bishop, D. Spiegelhalter, and J. Winn, "Vibes: A variational inference engine for bayesian networks," in *NIPS 2002*, 2002.

[15] D. Shutin and B. H. Fleury, "Sparse variational bayesian sage algorithm with application to the estimation of multipath wireless channels," *IEEE Transactions on Signal Processing*, vol. 59, pp. 3609–3623, Aug 2011.

[16] V. P. Oikonomou and D. I. Fotiadis, "A bayesian approach for the estimation of ar coefficients from noisy biomedical data," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3270–3273, Aug 2007.

[17] C. E. Shannon, "A mathematical theory of communication," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, pp. 3–55, Jan. 2001.

[18] R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics, Vol. II: The New Millennium Edition: Mainly Electromagnetism and Matter*. Feynman Lectures on Physics, Basic Books, 2011.

[19] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[20] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*. 2003.

[21] J. Ramirez, J. C. Segura, C. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Processing Letters*, vol. 12, pp. 689–692, Oct 2005.

[22] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Audio, Speech and Language Processing*, pp. 504–512, 2001.

[23] M. Kuropatwinski and W. B. Kleijn, "Estimation of the excitation variances of speech and noise ar-models for enhanced speech coding," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 1, pp. 669–672 vol.1, 2001.

[24] T. Quatieri, *Discrete-time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ, USA: Prentice Hall Press, first ed., 2001.

[25] W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*. New York, NY, USA: Elsevier Science Inc., 1995.

[26] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, pp. 84–95, 1980.

[27] S. Lloyd, "Least squares quantization in pcm," *IEEE Trans. Inf. Theor.*, vol. 28, pp. 129–137, Sept. 2006.

[28] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. No. 159 in The Kluwer International Series in Engineering and Computer Science, Kluwer, 1992.

[29] R. M. Gray, "Toeplitz and circulant matrices: A review," *Commun. Inf. Theory*, vol. 2, pp. 155–239, Aug. 2005.

[30] M. Kuropatwinski and W. B. Kleijn, "Estimation of the short-term predictor parameters of speech under noisy conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1645–1655, Sept 2006.

[31] R. Gray, A. Buzo, A. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 367–376, Aug 1980.

[32] K. Paliwal and A. Basu, "A speech enhancement method based on kalman filtering," in *ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, pp. 177–180, Apr 1987.

[33] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Transactions on Signal Processing*, vol. 39, pp. 1732–1742, Aug 1991.

[34] B.-G. Lee, K. Y. Lee, and S. Ann, "An em-based approach for parameter enhancement with an application to speech signals," *Signal Processing*, vol. 46, no. 1, pp. 1 – 14, 1995.

[35] J. Stahl and P. Mowlaee, "Iterative harmonic speech enhancement," in *Speech Communication; 12. ITG Symposium*, pp. 1–5, Oct 2016.

[36] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society, series B*, vol. 39, no. 1, pp. 1–38, 1977.

[37] R. E. Kalman, "A new approach to linear filtering and prediction problems," *ASME Journal of Basic Engineering*, 1960.

[38] C. L. Nikias and A. P. Petropulu, *Higher-order spectra analysis : a nonlinear signal processing framework / Chrysostomos L. Nikias, Athina P. Petropulu.* PTR Prentice Hall Englewood Cliffs, N.J, 1993.

[39] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin, "Variational bayesian inference for multichannel dereverberation and noise reduction," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, pp. 1320–1335, Aug. 2014.

[40] H. Permuter, J. Francos, and I. H. Jermyn, "Gaussian mixture models of texture and colour for image database retrieval," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 3, pp. III–569–72 vol.3, April 2003.

[41] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72–83, Jan 1995.

[42] J. Hao, T. W. Lee, and T. J. Sejnowski, "Speech enhancement using gaussian scale mixture models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1127–1136, Aug 2010.

[43] P. Mowlaee, R. Saeidi, and R. Martin, "Model-driven speech enhancement for multisource reverberant environment (signal separation evaluation campaign (sisec) 2011)," in *Latent Variable Analysis and Signal Separation* (F. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, eds.), (Berlin, Heidelberg), pp. 454–461, Springer Berlin Heidelberg, 2012.

[44] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes.* Boston: McGraw Hill, fourth ed., 2002.

[45] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of the Acoustics, Speech, and Signal Processing, 200. On IEEE International Conference - Volume 02*, ICASSP '01, (Washington, DC, USA), pp. 749–752, IEEE Computer Society, 2001.

[46] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time x2013;frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2125–2136, Sept 2011.

[47] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic phonetic continuous speech corpus cdrom," 1993.

[48] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems.," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[49] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, pp. 70–73, Jun 1967.

[50] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, pp. 443–445, Apr 1985.

[51] P. Mowlaee and C. Nachbar, "Speaker dependent speech enhancement using sinusoidal model," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 80–84, Sept 2014.