Speech Enhancement in the Modulation Domain

Yu Wang

Communications and Signal Processing Group

Department of Electrical and Electronic Engineering

Imperial College London

This thesis is submitted for the degree of

Doctor of Philosophy of

Imperial College London

August 2015

Copyright Declaration

Statement of Originality

I hereby certify that this thesis is the outcome of the research conducted by myself under the supervision from Mike Brookes in the Department of Electrical and Electronic Engineering at Imperial College London. Any work that has been previously published and included in this thesis has been fully acknowledged in accordance with the standard referencing practices of this discipline. I declare that this thesis has not been submitted for any degree at any other University or Institution.

Abstract

The goal of a speech enhancement algorithm is to reduce or eliminate background noise without distorting the speech signal. Although speech enhancement is important for practical scenarios, it is a difficult task especially when the noisy speech signal is only available from a single channel. Although many single-channel speech algorithms have been proposed that can improve the Signal-to-Noise Ratio (SNR) of the noisy speech, in some cases dramatically, they also introduce speech distortion and spurious tonal artefacts known as musical noise.

There has been evidence, both physiological and psychoacoustic, to support the significance of the modulation domain, i.e. the temporal modulation of the acoustic spectral components, to speech enhancement. In this thesis three methods for implementing single-channel speech enhancement in the modulation domain have been proposed. The goal in all three cases is to take advantage of prior knowledge about the temporal modulation of short-time spectral amplitudes. The first method is to post-process the output of a conventional single-channel speech enhancement algorithm using a modulation domain Kalman filter. The second method performs enhancement directly in the modulation domain based on the assumption that the temporal sequence of spectral amplitudes within each frequency bin lies within a low dimensional subspace. The third method uses a modulation-domain Kalman filter to perform enhancement using two alternative distribution families for the speech and noise amplitude prior distributions. The performance of the proposed enhancement algorithms is assessed by measuring the SNR and speech quality (using the Perceptual Evaluation of Speech Quality (PESQ) metric) of the enhanced speech. It is found that, for a range of noise types, the proposed algorithms give consistent improvements in both metrics.

Acknowledgements

First and foremost, I would like to thank my PhD supervisor, Mr Mike Brookes, for his guidance and support throughout my PhD study at Imperial College London. His expertise, insight and encouragement have always been very helpful and the weekly meetings with him have always been a source of the new ideas for my research work. Also, his patience at teaching and reviewing benefits students significantly especially for non-native English speakers like me. It has been a privilege and memorable experience to work with him.

Second, a note of thanks goes to my colleagues in the Speech and Audio Processing group, in no particular order: Sira Gonzalez, Feilicia Lim, Richard Stanton, Alastair Moore, Hamza Javed and Leo Lightburn, with whom I spent most of the enjoyable time at Imperial. I will indeed miss the time working with you. I also want to add special thanks to my internship supervisor at Nuance Communications, Dr Dushyant Sharma, for his advice and assistance during my research in the machine learning field at Nuance.

Last but not least, I would like to thank my family, particularly my parents, Minsheng and Xiufen, and also my wife, Wenshan. Without their unending love and support this research work would not be possible to come to fruition.

Table of Contents

List of Figures

List of Acronyms

GMM Gaussian Mixture Model. An approximation to an arbitrary probability density function that consists of a weighted sum of Gaussian distributions 70

KFMD Modulation Domain Kalman filter post-processor 60

KFGM Kalman filter post-processor with a GMM noise model 75

KLT Karhunen-Loéve Transform 36

KMMSE Kalman filter based MMSE estimator 114

KMMSEI Intermediate KMMSE 114

LMS Least Mean Squares adaptive filter 3

LPC Linear Predictive Coding. An autoregressive model of speech production. 14

LTASS Long Term Average Speech Spectrum 11

LTANS Long Term Average Noise Spectrum 16

LTASMS Long Term Average Speech Modulation Spectrum 12

LTANMS Long Term Average Noise Modulation Spectrum 20

logMMSE log-amplitude MMSE 102

MAP Maximum a Posteriori 43

MDKF Modulation Domain Kalman filter that assumes white noise . . 113

MDKFR Modulation Domain Kalman filter based on a Gaussring model 135

List of Symbols

d_s	dc component of speech amplitudes
$\tilde{e}_n$	prediction residual signal of speech
$\breve{e}_n$	prediction residual power of speech
$h(t)$	acoustic domain window function
$\check{h}_n$	modulation domain window function
k	acoustic frequency index
m	shape parameter of the Nakagami-m distribution
n	acoustic frame index
p	LPC order of speech
q	LPC order of noise
$s(t)$	time-domain clean speech
t	time sample index
$w(t)$	time-domain noise
$z(t)$	time-domain noisy speech
$A_{n,k}$	spectral amplitude of clean speech
$F_{n,k}$	spectral amplitude of noise
G_H	DC gain of LPC synthesis filter
$H(z)$	LPC synthesis filter
J	number of GMM mixtures
L	modulation frame length

M	acoustic frame increment
Q	modulation frame increment
$R_{n,k}$	spectral amplitude of noisy speech
$S_{n,k}$	Complex STFT coefficients of clean speech
$S_l(n,k)$	modulation frame of clean speech
T	acoustic frame length
$W_{n,k}$	Complex STFT coefficients of noise
$W_l(n,k)$	modulation frame of noise
$\widetilde{W}_{n,k}$	Complex STFT coefficients of white noise
$Y_{n,k}$	Complex STFT coefficients of MMSE enhanced speech
$Z_{n,k}$	Complex STFT coefficients of noisy speech
$Z_l(n,k)$	modulation frame of noisy speech
$\mathbf{a}_c(k)$	autocorrelation coefficients vector of noise
$\tilde{\mathbf{b}}_n$	speech LPC coefficients vector
$\breve{\mathbf{b}}_n$	noise LPC coefficients vector
$\mathbf{k}_n$	Kalman gain
$\mathbf{g}$	autocorrelation coefficients vector of speech
$\mathbf{o}$	vector of ones
$\tilde{\mathbf{s}}_n$	state vector of speech
$\breve{\mathbf{s}}_n$	state vector of noise
$\mathbf{s}_n$	augmented state vector
$\mathbf{s}_l$	modulation-domain speech vector
$\mathbf{w}_l$	modulation-domain noise vector
$\mathbf{z}_l$	modulation-domain noisy speech vector
$\widetilde{\mathbf{A}}_n$	speech transition matrix
$\breve{\mathbf{A}}_n$	noise transition matrix
$\widetilde{\mathbf{\Sigma}}_n$	error covariance matrix of the speech state vector

$\breve{\mathbf{\Sigma}}_n$	error covariance matrix of the noise state vector
$\mathbf{\Sigma}_n$	error covariance matrix of the augmented state vector
$\mathbf{Q}_n$	covariance matrix of prediction residual signal
$\mathbf{R}$	autocorrelation matrix of speech
$\mathbf{R}_S$	covariance matrix of modulation-domain speech vector
$\mathbf{R}_W$	covariance matrix of modulation-domain noise vector
$\mathbf{R}_Z$	covariance matrix of modulation-domain noisy speech vector
$\mathbf{H}_l$	subspace estimator of clean speech
$\mathbf{U}$	eigenvector matrix of whitened noisy speech
$\mathbf{P}$	diagonal matrix consisting of eigenvalues
β_n	scale parameter of the Gamma distribution
$\epsilon^{(j)}$	weight of Gaussian mixtures
η	Lagrange multiplier
$\theta_{n,k}$	phase spectrum of noisy speech
$\phi_{n,k}$	phase spectrum of clean speech
κ	forgetting factor for updating GMM parameters
γ_n	shape parameter of the Gamma distribution
$\xi_{n,k}$	a priori SNR
$\zeta_{n,k}$	a posteriori SNR
$\pi^{(j)}$	responsibility of each GMM mixture
$\tilde{\sigma}_n^2$	prediction residual power of speech
$\breve{\sigma}_n^2$	prediction residual power of noise
$\nu_{n,k}^2$	power spectrum of colored noise
ν_w^2	power spectrum of white noise
σ_w^2	variance of white noise in time domain
Λ	eigenvalues of covariance matrix of speech
Ω	spread parameter of the Nakagami-m distribution

1. Introduction

1.1. Speech Enhancement

In practical situations, clean speech signals are often contaminated by unwanted noise from the surrounding environment or communication channels. As a result, speech enhancement is often needed, the goal of which is to remove of the noise and improve the perceptual quality of the speech signal. There are different types of noise, which include additive acoustic noise, convolution noise, and transcoding noise [1]. Additive noise that is uncorrelated with the clean speech signal in either the acoustic or electronic domain. Its perceived effect is to degrade the quality and intelligibility, and may, in extreme cases, completely mask the clean speech signal. Convolution noise is perceived as reverberation and poor spectral balance. Reverberation is normally introduced by acoustic reflections and can seriously degrade intelligibility. This type of noise differs from the additive noise in that it is strongly correlated with the clean speech signal. Transcoding noise normally arises from amplitude limiting or clipping in the microphone, amplifier or CODEC and it is perceived as severe distortion that varies with the amplitude of the speech signal. In this thesis the removal of additive acoustic noise will be concerned. Speech enhancement methods may be divided into two types. The first one is single channel methods where the signal from unique acquisition channel is available. The second

type is multi channel methods [2] where multi speech signals can be obtained from a number of microphones, and the noise reduction can be achieved making use of the information (e.g. noise reference, phase alignment) provided from each of the microphones and thus the Signal-to-Noise Ratio (SNR) can be improved. Although multiple channel methods often yield better performance than single channel methods, they also introduce additional costs, such as power usage, computational complexity and requirement of size. As a result, single channel methods are necessary in many devices where multi microphone methods cannot be applied, such as mobile phones, hearing aids and cochlear implant devices, most of which have only a single microphone due to the limit on the location and size of the devices. In this thesis, on the single channel speech enhancement task will be focused.

Over the past three decades, numerous single channel speech enhancement algorithms have been presented [3]. The main issues with the single channel speech enhancement problem include: 1) need to attenuate noise without introducing artifacts or distorting the speech; 2) need to distinguish between speech and noise on the basis of their differing characteristics; 3) varying acoustic noise arising from many sources, such as car engine and factory machine, and so far no universal model which represents all possible noises well has been proposed.

1.2. Enhancement Domains

Speech enhancement can be performed in several alternative domains. The following sections define these alternative domains and describe illustrative enhancement algorithms. A more complete review of speech enhancement algorithms relevant to this thesis is given in Chapter 2.

1.2.1. Time domain

In the time domain, enhancement is normally achieved by making the use of static or adaptive filtering techniques. Two types of the well known adaptation algorithms are the Least Mean Squares (LMS) algorithm, or more commonly, the Normalized Least Mean Squares (NLMS) gradient descent algorithm [4], and the Recursive Least Squares (RLS) algorithm. The adaptive filtering for single channel speech enhancement was introduced in [5] with two applications. The diagram of the algorithm is given in Figure 1.1, in which t denotes the discrete time index. The noisy speech signal, $z(t)$, is firstly delayed by D samples, where D is an integer, and is processed by an LMS adaptive filter to give the signal $y(t)$, which is then subtracted from $z(t)$ to produce the error signal $e(t)$. The output of the filter is generated by a mixture of $e(t)$ and $y(t)$ which depends on whether the periodic signal components should be suppressed or enhanced. When $\alpha = 1$, it brings the first application that the filter can be used for removing periodic noise from a broadband speech signal [5], because a fixed delay is inserted in the input of the adaptive filter, which is obtained directly from the original input. In this case, the delay needs to be long enough so that $z(t)$ and $z(t - D)$ are uncorrelated. When $\alpha = 0$ (e.g. $\hat{s}(t) = y(t)$), the function of the filter turns into the reverse of the first application and it aims to remove broadband noise from a periodic signal. Because both periodic and broadband components are often present in both speech and noise, it is important to chose the parameters of the filtering properly to enhance the wanted components.

1.2.2. Time-frequency domain

The enhancement can also be applied in the Time-Frequency (TF) domain. In this domain, speech samples are divided into frames, which will be referred to as

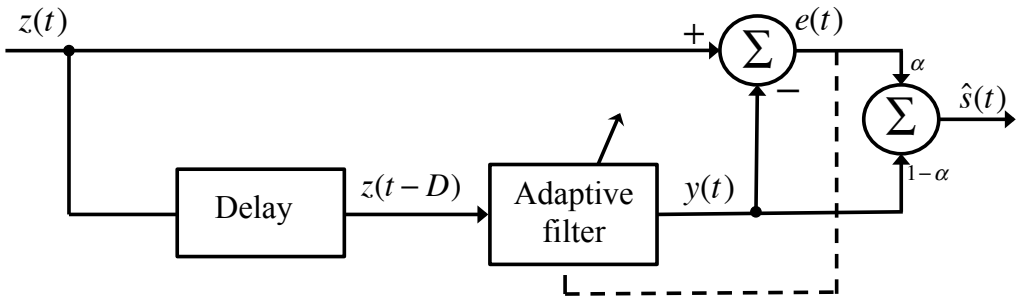

Figure 1.1.: Adaptive filtering for enhancement

acoustic frames in order to distinguish them from the *modulation frames* that will be introduced in Section 1.2.3. The diagram of TF domain speech enhancement algorithms is given in Figure 1.2. Let $s(t)$ and $w(t)$ denote the speech and noise in the time domain, respectively. The noisy speech $z(t)$ is given by

$$z(t) = s(t) + w(t) \tag{1.1}$$

A Short Time Fourier Transform (STFT) is firstly applied to the noisy speech $z(t)$, which is defined as

$$Z_{n,k} = \sum_{t=0}^{T-1} z(nM + t)h(t)e^{-2\pi j \frac{tk}{T}} \tag{1.2}$$

where n and k denote the time frames and frequency bins respectively. T is the acoustic frame length in samples and $M \leq T$ is the time increment between successive frames. The frame length is a compromise between frequency and time resolution and is typically chosen in the range $10 - 30$ ms therefore T is in the range 80 to 240 samples when the sampling frequency in 8000 Hz. $h(t)$ is the window (e.g. Hamming window) used to segment the time-domain speech into short-time frames. The speech and noise can be transformed into the STFT domain in the same way to obtain the STFT coefficients $S_{n,k}$ and $W_{n,k}$, respectively. The general framework of TF processing applies a real-valued TF gain function with the aim

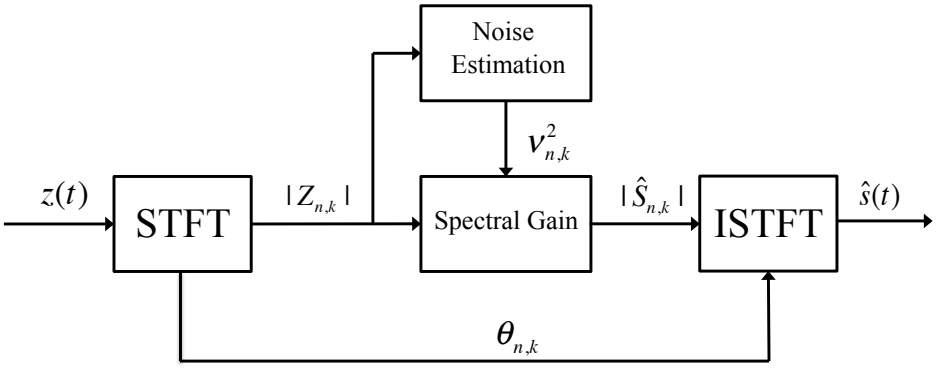

Figure 1.2.: Diagram of time-frequency domain speech enhancement

of suppressing noise-dominated TF regions while preserving the speech-dominated TF regions. From the STFT, the noisy amplitude spectrum $|Z_{n,k}|$ and phase spectrum $\theta_{n,k} = \angle Z_{n,k}$ for frame n is obtained. Since the phase information is widely considered to be unimportant in the perception of speech signals [6], only the noisy amplitude spectrum is processed by a spectral attenuation gain which is derived under assumptions on the statistical characteristics of the time-frequency signals of speech and noise [7, 8]. The calculation of the gain function typically depends on the noise power spectrum $\nu^2_{n,k} = \mathrm{E}(|W_{n,k}|^2)$, where $\mathrm{E}(\cdot)$ is the expectation operator. Noise power can be estimated using the methods reviewed in Section 2.2. After the estimated amplitude spectrogram of the clean speech, $\hat{S}_{n,k}$, is obtained, which is combined with the phase spectrum of the noisy speech, $\theta_{n,k}$. The inverse STFT (ISTFT) is then applied to give the enhanced speech signal $\hat{s}(t)$. The reconstruction properties can be controlled by the choice of the window and the ratio M/T. It is found that a three-quarters overlap ($M = T/4$) is needed to avoid aliasing in the spectral coefficients when a Hamming window is used [9, 10].

The reason why the TF domain processing works is that speech is sparse, as shown in the left spectrogram in Figure 1.3 which obtained from a sentence in the TIMIT

database [11]. Although the TF enhancement can dramatically improve the SNR of the noisy speech, it usually introduces "musical noise" artefacts, which can be illustrated in the middle and right spectrograms in Figure 1.3. In the middle spectrogram, the clean speech is corrupted by factory noise at -5 dB SNR and the right spectrogram shows the spectrogram of the enhanced speech using a Minimum Mean Squared Error (MMSE) based TF domain enhancement algorithm in [7] (for details see Section 2.4). It can be seen that although the speech enhancement has greatly reduced the level of the noise, isolated spectral components of the noise remain throughout the spectrogram. This is due to the fact that, after the TF domain processing the spectrogram now consists of a succession of randomly spaced spectral peaks corresponding to the maxima of the original spectrogram. Thus, the residual noise consists of sinusoidal components with random frequencies which exist in between each short-time frame. They manifest as brief tones in the enhanced speech and are known as "musical noise" [12]. This problem will be discussed in more detail in Chapter 3.

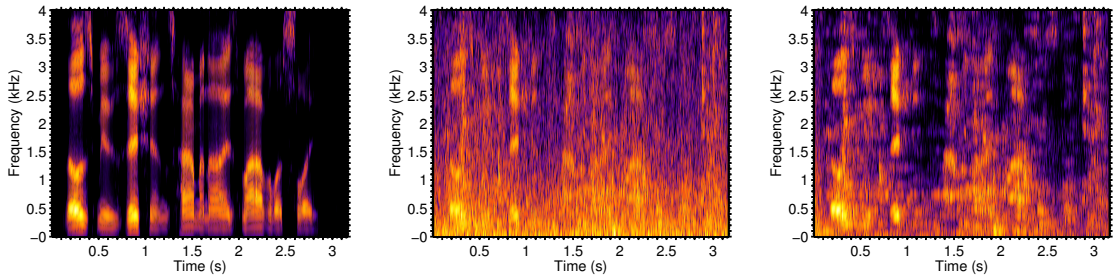

Figure 1.3.: Spectrogram of clean speech (left), noisy speech (center) and enhanced speech (right), where the speech signal is corrupted by factory noise at -5 dB and the speech enhancement uses the algorithm from [7].

1.2.3. Modulation domain

The diagram of modulation-domain processing is given in Figure 1.4. The first step is to segment the temporal sequence of spectral amplitudes into *modulation frames*. For speech enhancement, the noisy spectral amplitudes envelope of each frequency band $|Z_{n,k}|$ is segmented into overlapped modulation frames $Z_l(n,k)$ of length L with a frame increment Q multiplied by a window function, which is

$$Z_l(n,k) = \check{h}_n |Z_{lQ+n,k}| \;\; n = 0 \cdots L-1 \tag{1.3}$$

where $|\cdot|$ denotes the absolute value of a complex number, l is the modulation frame index and $\check{h}_n$ is the window applied to segment the envelope of the speech STFT amplitudes. In this thesis, the acoustic frame index, n, and acoustic frequency index, k, will be put in the subscript to save space. A graph showing about the process to obtain $Z_l(n,k)$ is shown in Figure 1.5. Because there are L acoustic frames forming one modulation frames and one acoustic frame is constructed by T time-domain speech samples, one modulation frame is constructed by TL time-domain samples. The acoustic frame increment M determine the sampling frequency for the modulation-domain signal. If the time-domain sampling frequency is 8000 Hz, then the modulation sampling frequency is $\frac{8000}{M}$ Hz. An estimator is then applied to each modulation frame of noisy speech $Z_l(n,k)$ to estimate the modulation frames of clean speech $\hat{S}_l(n,k)$, which are then used to give the spectral envelopes $|\hat{S}_{n,k}|$ by overlap-add. The time-domain estimated clean speech $\hat{s}(t)$ can then be obtained by combining $|\hat{S}_{n,k}|$ with the phase spectrum $\theta_{n,k}$ and applying an ISTFT. The enhancement processing can be applied either directly to the amplitude envelope, $Z_l(n,k)$, or to the amplitude spectrum of each modulation frame (known as either the modulation spectrum or the amplitude modulation spectrum) [13, 14, 15]. The

essential difference between time-frequency domain processing and modulation domain processing is that for the later, the long-term correlation between the samples of time-frequency amplitudes within each frequency bin is considered in the development of models or techniques.

Speech modulation is closely related to speech intelligibility. For instance, the Speech Transmission Index (STI) measure, which is designed to predict the intelligibility of both linear and nonlinear distortions [16], is based on the effect on the modulation depth, which is defined as the ratio of the modulation singal amplitude to the carrier signal amplitude, within several frequency bands at the output of the communication channel. STI has been proven to be successful in predicting intelligibility for a variety of practical situations such as noisy and reverberant environment. As an extension of the STI measure, the Short-Time Objective Intelligibility Measure (STOI) measure, proposed in [17], calculates the sample correlation coefficient between the spectral modulations of the clean speech and that of the noisy speech as the intermediate intelligibility measure, which shows higher correlation with speech intelligibility than the STI measure for TF-weighed speech.

The modulation frequency components represent the rate of change of human speech production which is caused by the dynamics of the glottal source and those of the vocal tract, since the airflow generated by the lungs is modulated by this overall dynamics, the modulation components will convey the information which can separate speech from other interference such as noise or reverberation. Most of the modulation energy of speech is distributed at modulation frequencies 4 to 16 Hz while other modulation frequencies has most of the energy for noise [18, 19].

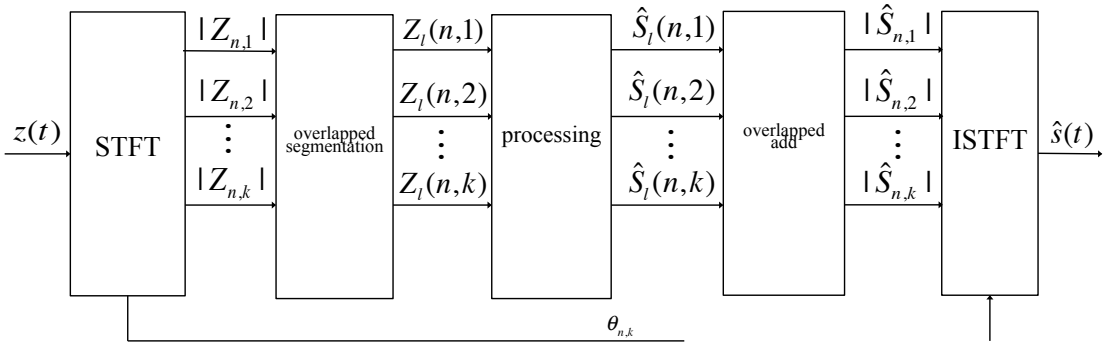

Figure 1.4.: Diagram of modulation domain processing

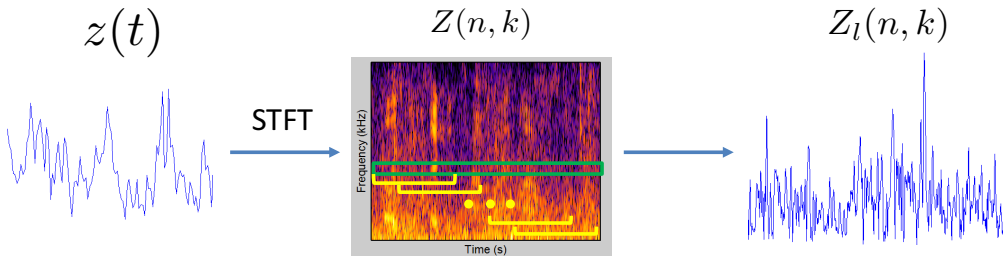

Figure 1.5.: Steps to obtain modulation frames $Z_l(n,k)$

1.3. Goal of Research

Based on the observation on the importance of the modulation of the spectral amplitudes of the speech signal and noise, the main research aim is to develop single-channel speech enhancement algorithms for speech corrupted by acoustic additive noise using the modulation-domain characteristics of speech and noise signals.

1.4. Speech and Noise Databases

There have been a number of publicly or commercially available speech and noise databases which may be suitable for evaluating speech enhancement algorithms.

This section gives a brief overview of the speech and noise databases which will be used to assess the performance of different algorithms in this thesis, and the acoustic and modulation spectral characteristics of typical speech and noise will be shown. The long-term and short-term acoustic spectrograms and modulation spectrograms of speech will be shown and typical types of noises are given as follows.

1.4.1. Speech database

1.4.1.1. TIMIT

The TIMIT database was designed jointly by the Massachusetts Institute of Technology (MIT), SRI International (SRI) and Texas Instruments, Inc. (TI). The TIMIT database consists of broadband recordings of 630 speakers of eight major dialects of American English, each of whom speaks 10 sentences lasting a few seconds each and the length of the entire database is about 5.4 hours [11]. The database is recorded using a microphone at 16 kHz rate with a 16 bits sample resolution. All the recordings are manually segmented at the phone level. TIMIT has been widely used in speech related research for more than two decades. For evaluating the speech enhancement algorithms proposed in this thesis, the core test set of the TIMIT database will be used which contains 16 male and 8 female speakers each reading 8 sentences for a total of 192 sentences all with distinct texts. This test set is the abridged version of the complete TIMIT test set which consists of 1344 sentences from 168 speakers. Also, in order to optimize the parameters of the algorithms, a development set is formed which consists of 200 speech sentences randomly selected from the test set of the TIMIT database and does not have any overlap with the core test set. The speech sentences in the development set are corrupted by white noise, car noise, factory noise, F16 noise and babble noise at SNRs between -10 and 15 dB at a

interval of 5 dB. All the the speech sentences used in this thesis are downsampled to 8000 Hz.

1.4.1.2. LTASS and spectrogram of speech

The Long Term Average Speech Spectrum (LTASS) [20] has a characteristic shape that is often used as a model for the clean speech spectrum and has been used in a wide range of speech processing algorithms, such as blind channel identification [21]. The LTASS of speech signal can be estimated by the average smoothed STFT power spectrum of all the acoustic frames that are mostly active. The LTASS averaged over 65 seconds of speech sentences from the TIMIT database is given in Figure 1.6 and the spectrogram of one speech sentence is given in Figure 1.7.

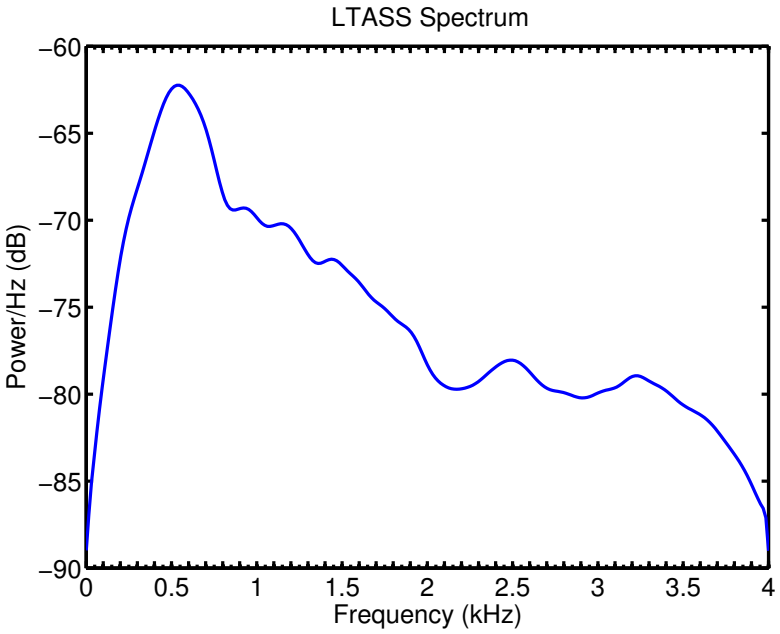

Figure 1.6.: LTASS of speech from the TIMIT database, which is obtained by averaging over about 65 seconds of speech sentences.

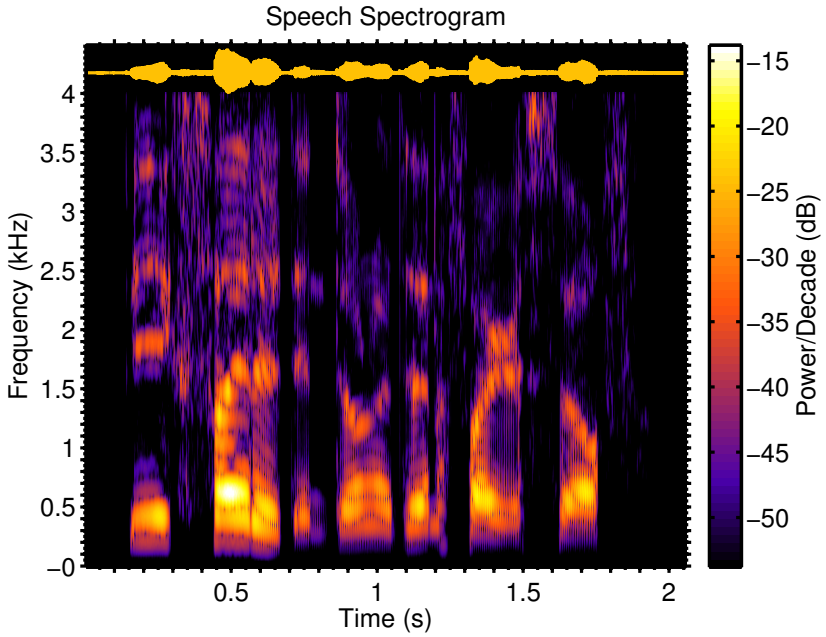

Figure 1.7.: Spectrogram and the time domain signal of one speech sentence from the TIMIT database.

1.4.1.3. LTASMS and spectrogram of speech

Compared with the acoustic spectral characteristics of speech, the Long Term Average Speech Modulation Spectrum (LTASMS) and the corresponding short-time modulation spectrogram are given in Figure 1.8 and 1.9, respectively. The modulation spectra are taken for the acoustic frames sequence at acoustic frequency of 500 Hz. As shown in the modulation spectra, most of the speech modulation energy concentrates at low modulation frequencies, which is consistent with the observation described in Section 1.2.3.

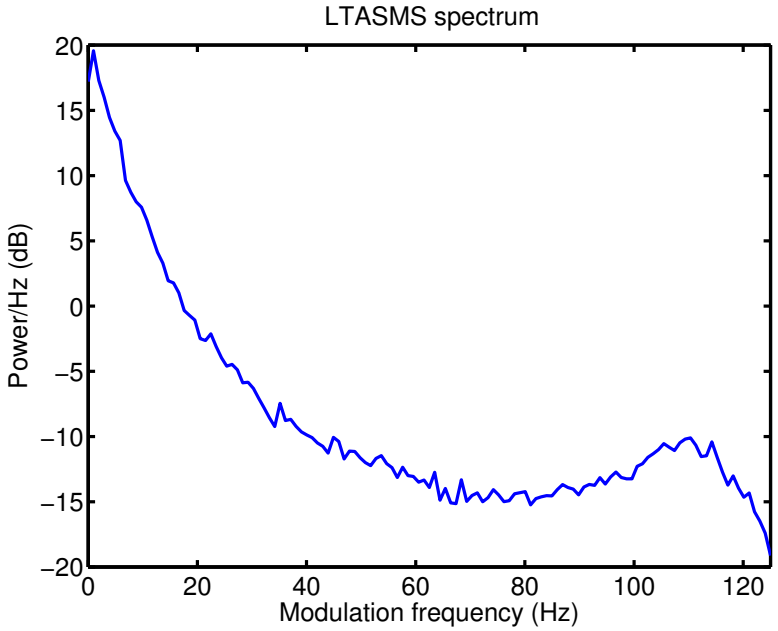

Figure 1.8.: LTASMS of one acoustic frequency bin (500 Hz), which is obtained by averaging over about 65 seconds of speech sentences.

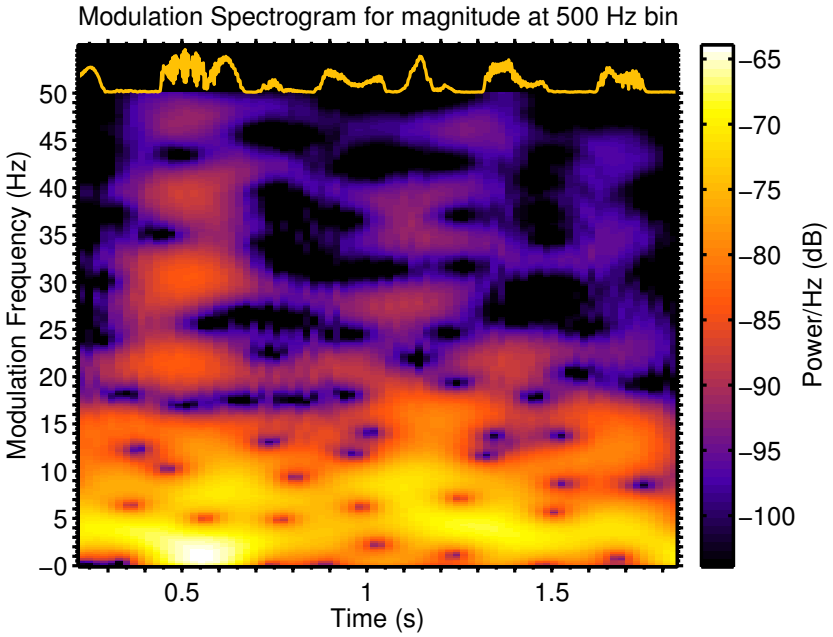

Figure 1.9.: Modulation spectrum of one acoustic frequency bin (500 Hz), the speech sentence is from the TIMIT database.

1.4.1.4. Modulation domain LPC of speech

Linear Predictive Coding (LPC) model has been widely used in the speech analysis and synthesis fields [22]. The basis of the LPC model is that the speech signal is generated by a low-order autogressive process and therefore its covariance matrix is rank-deficient [23]. The conventional LPC model is applied on the time-domain speech signal and because the signal is non-stationary, it is normally segmented into short-time frames before the LPC analysis. In the following chapters, the LPC model will be applied in the modulation domain when using a modulation-domain Kalman filter for speech enhancement. To validate that the speech modulation of each frequency bin can be predicted using a LPC mode, the prediction gain of different LPC order is shown in Figure 1.10. The prediction gain is defined as [24]

$$G_p \triangleq \frac{\mathrm{E}\left(|S_{n,k}|^2\right)}{\mathrm{E}\left(\left(|S_{n,k}| - |\widehat{S}_{n,k}|\right)^2\right)} \tag{1.4}$$

where $\mathrm{E}(\cdot)$ is the expectation operator. $|\widehat{S}_{n,k}|$ is the predicted amplitude. The expectation is taken over all acoustic frames, n, at frequency bin k, and Figure 1.10 was formed using 100 speech sentences from the core test set of the TIMIT dataset. The speech signals are segmented into acoustic frames of 32 ms with 4 ms increment. The LPC coefficients are estimated from modulation frame of 128 ms (thus there are 32 acoustic frames in one modulation frame). All the speech signals are downsampled from 16000 Hz to 8000 Hz. From Figure 1.10 it can be seen that, when the order of the modulation domain LPC model is ≥ 2, the prediction gain for most of the acoustic frequencies are larger than 10 dB. For the acoustic frequencies with most of the speech power ($500 - 1000$ Hz), the prediction gain is larger than 15 dB. In this thesis 3-order LPC models are used for a balance between the modeling capability and computational complexity. It is worth nothing that in the algorithms presented

in this thesis, a positive-valued floor has been applied to the speech amplitudes predicted by the modulation-domain LPC models by imposing the constraint

$$|\widehat{S}_{n,k}| = \max(|\widehat{S}_{n,k}|, 0.1|Z_{n,k}|)$$

The same floor is also imposed to the predicted noise amplitudes when a noise modulation-domain LPC model is also applied.

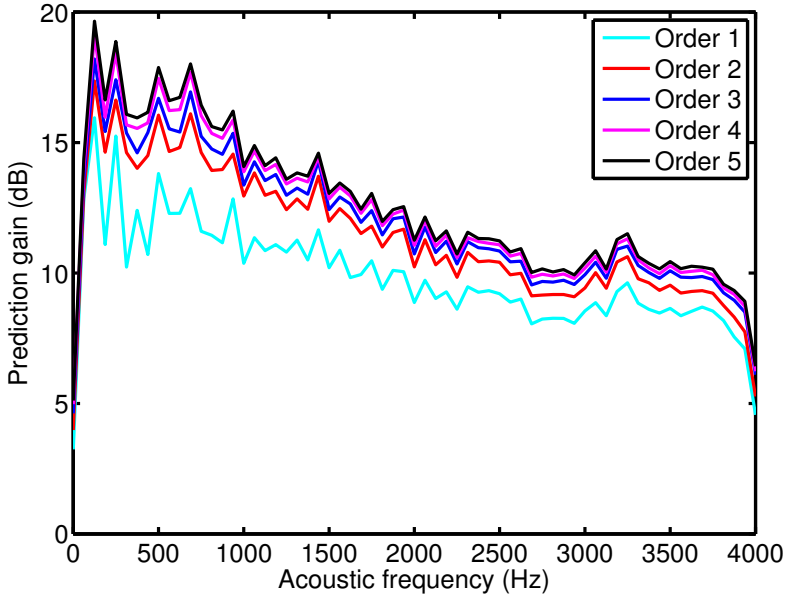

Figure 1.10.: Prediction gain of modulation-domain LPC model of different orders for speech. The speech power and prediction error power are averaged over all the acoustic frames of 100 speech sentences from TIMIT database.

The predictability of the modulation envelope of the speech signal is one of the primary motivations of the work in this thesis. Most existing enhancement algorithms do not take it into account explicitly although using a decision directed SNR estimate as in MMSE does implicitly assume correlation between adjacent acoustic frames of the clean speech.

1.4.2. Noise databases

1.4.2.1. RSG-10 noise database

The RSG-10 noise database is produced by the NATO Research Study Group on Speech Processing [25], which consists of 18 types of noises representative of military situations plus some more situations in addition to some civilian noises such as car and multitalker babble noise. The noises are recorded at 19.98 kHz with a 16 bits sample resolution. In this thesis, five of the noises from the RSG-10 database are primarily used: white noise, car noise, factory noise, F16 noise and babble noise and all the noises are downsampled to 8000 Hz.

1.4.2.2. ITU-T test signals

The ITU-T test signals are comprised of different test signals with different levels of complexity and designed for different types of applications, which includes fully artificial signal, speech-like signals and speech signals. In the fully artificial signals set and speech-like signals set there includes random noise (e.g. white noise, pink noise) and speech-like modulated noise, which consists of monaural noises (e.g. cafetaria noise, street noise). The nosies are recorded at 16 kHz with a 16 bits sample resolution. In this thesis, street noise from the ITU-T test signals is primarily used and it is downsampled to 8000 Hz.

1.4.2.3. LTANS and spectrogram of noise

The Long Term Average Noise Spectrum (LTANS) and the corresponding spectrograms for three different types of noises from the RSG-10 noise database and ITU-T test signals are given from Figure 1.11 to Figure 1.16. The mean of the frames of the noise signals are removed after the windowing and before the STFT applied. The

three noises have different spectral characteristics: the white noise has a constant power spectrum which does not depend on the frequency, while the power spectra of car and street noises is not stationary. The LTANS spectrum of shown in Figure 1.11 decreases at high and low frequencies because of the frame segmentation and the windowing. It is also worth noting that the intensity scale of spectrograms in this thesis is in Power/Decade rather than Power/Hz in which the power spectral density at a frequency f is multiplied by $\ln(10) \times f$. This pre-emphsis makes high frequency spectral components more visible in the spectrograms. Thus the intensity of the white noise in Figure 1.12 is increased at high frequencies. Most of the power of the car noise, as shown in Figure 1.14, concentrates at the low acoustic frequencies while the power of the street noise is more widely distributed over frequencies.

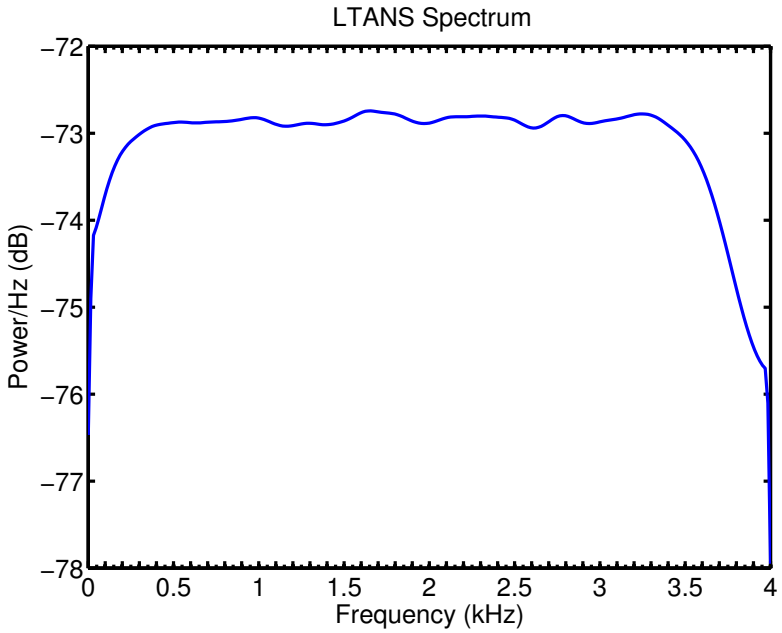

Figure 1.11.: LTANS of white noise, which is obtained by averaging over about 65 seconds of white noise signal.

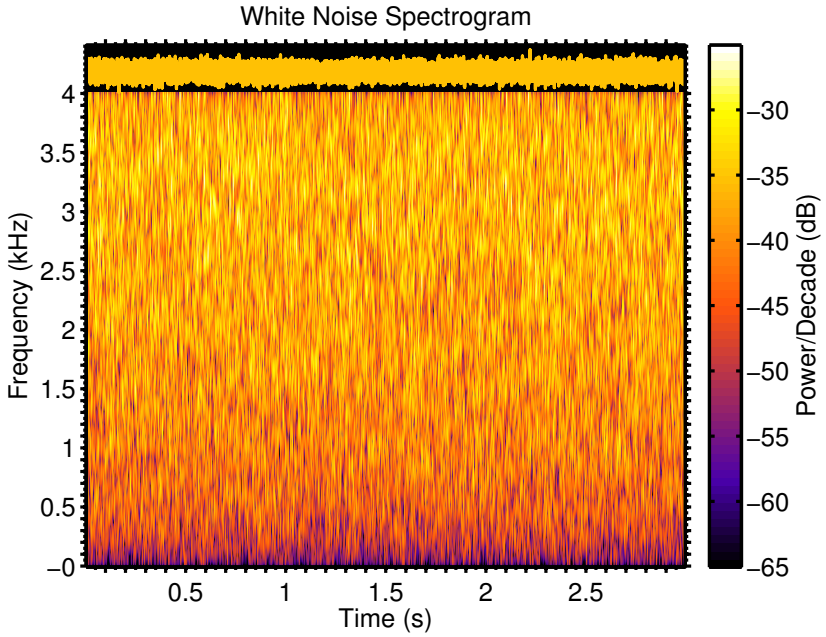

Figure 1.12.: Spectrogram and the time domain signal of white noise from RSG-10 noise database.

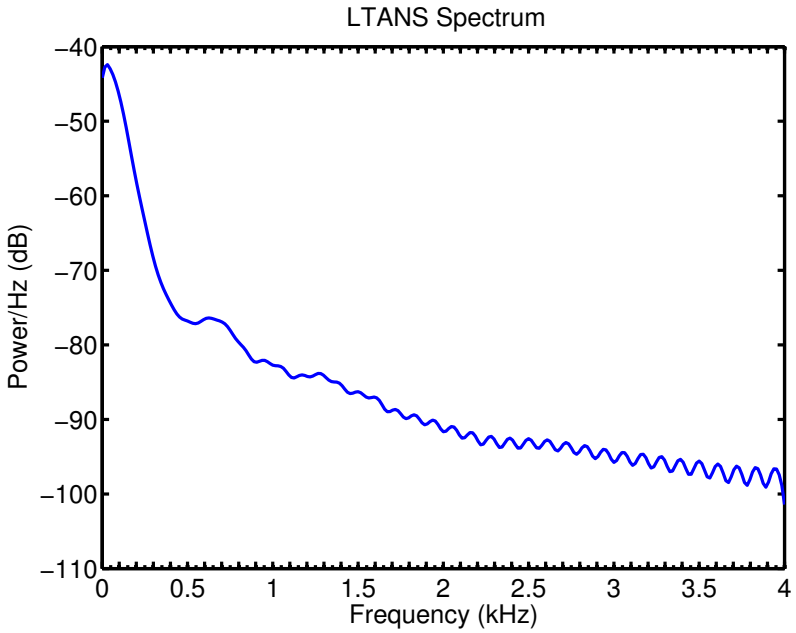

Figure 1.13.: LTANS of car noise from RSG-10 noise database, which is obtained by averaging over about 65 seconds of car noise signal.

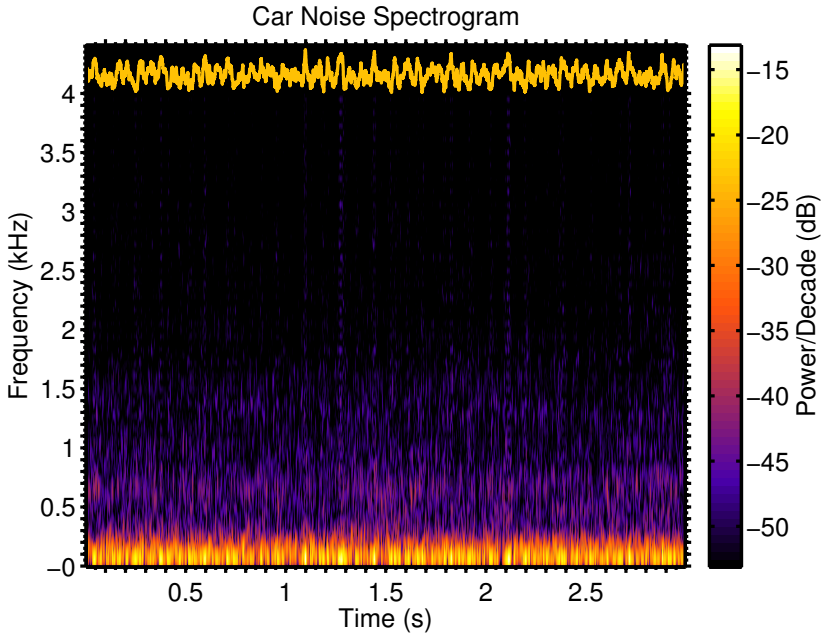

Figure 1.14.: Spectrogram and the time domain signal of car noise from RSG-10 noise database.

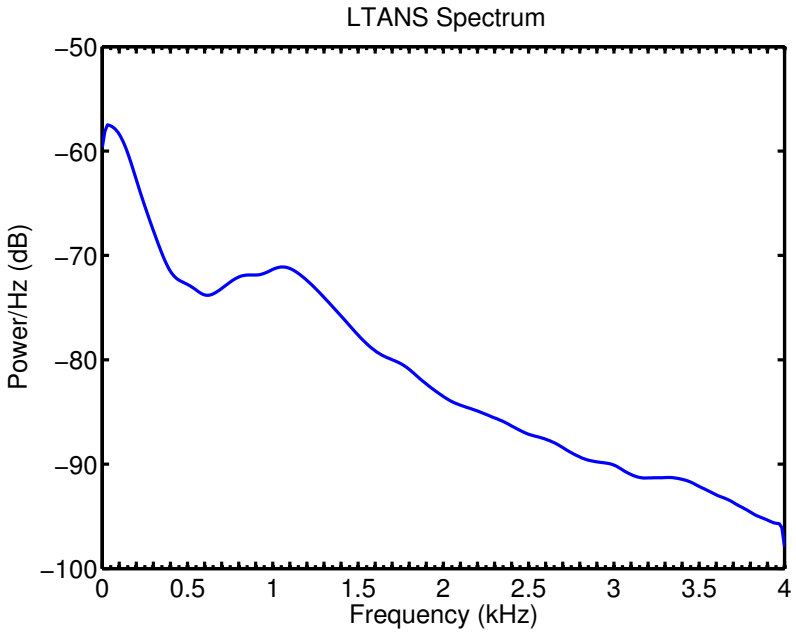

Figure 1.15.: LTANS of street noise from ITU-T test signal database, which is obtained by averaging over about 65 seconds of street noise signal.

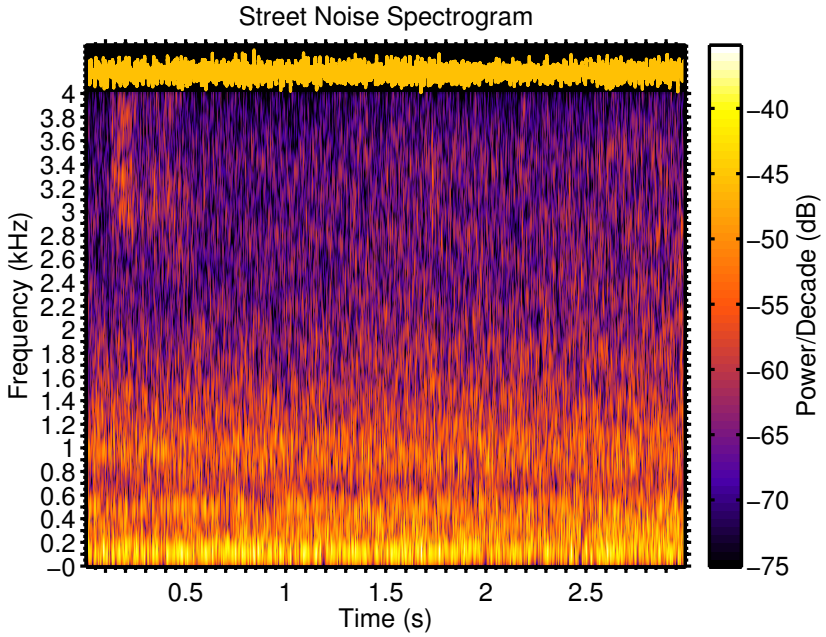

Figure 1.16.: Spectrogram and the time domain signal of street noise from ITU-T test signal database.

1.4.2.4. LTANMS and modulation spectrogram of noise

The Long Term Average Noise Modulation Spectrum (LTANMS) and the corresponding modulation spectrograms for the three different noises are shown from Figure 1.17 to Figure 1.22, which are calculated by taking the Fourier transform of the acoustic frames sequence at 500 Hz acoustic frequency. Compared to the modulation spectrograms of the speech, the modulation power of the noises are more widely distributed and more power contained at high modulation frequencies. The figures also show that the distribution of the modulation power of different types of noise are fairly consistent.

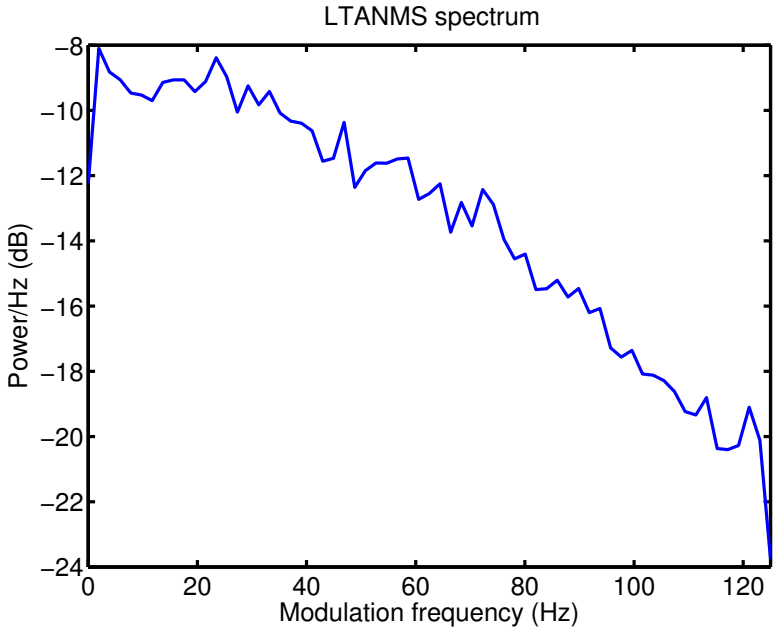

Figure 1.17.: LTANMS of white noise from RSG-10 noise database, which is obtained by averaging over about 65 seconds of white noise signal.

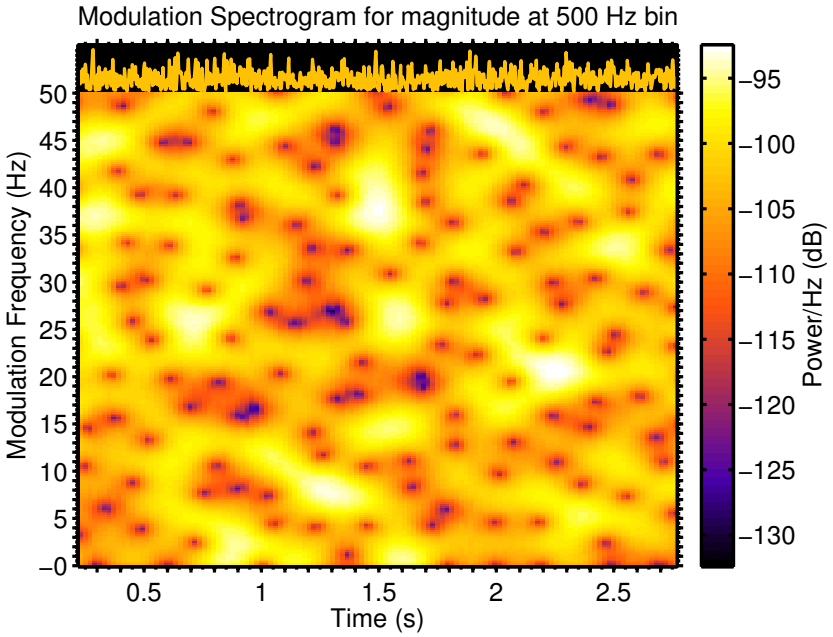

Figure 1.18.: Modulation spectrum of white noise from RSG-10 noise database.

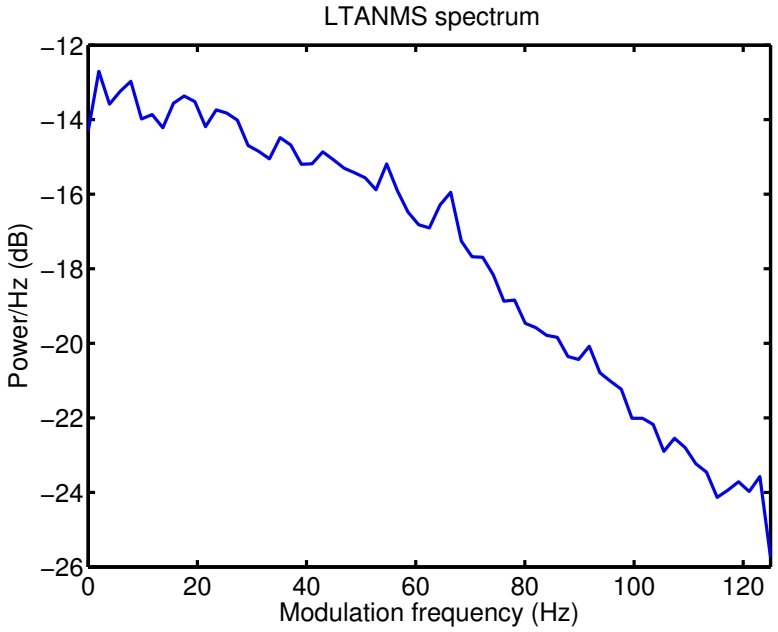

Figure 1.19.: LTANMS of car noise from RSG-10 noise database, which is obtained by averaging over about 65 seconds of car noise signal.

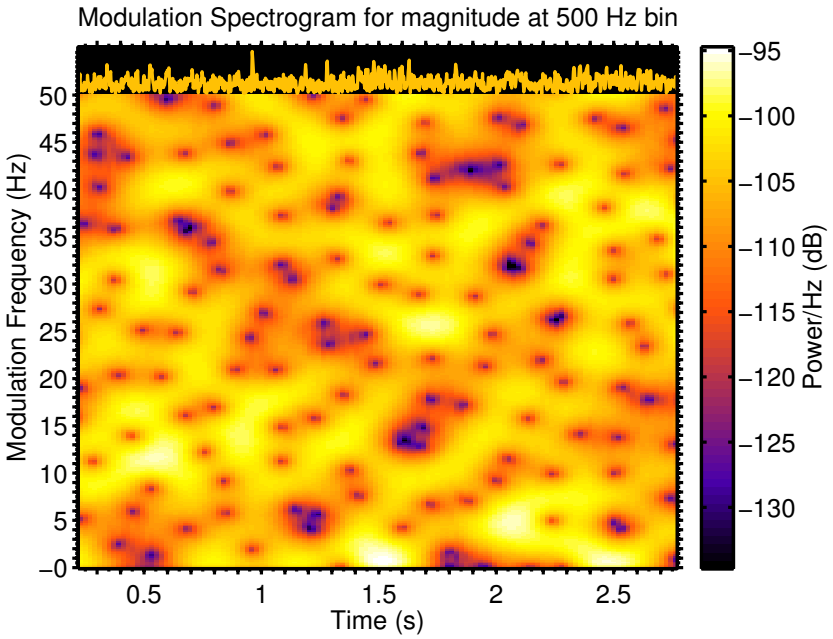

Figure 1.20.: Modulation spectrum of car noise from RSG-10 noise database.

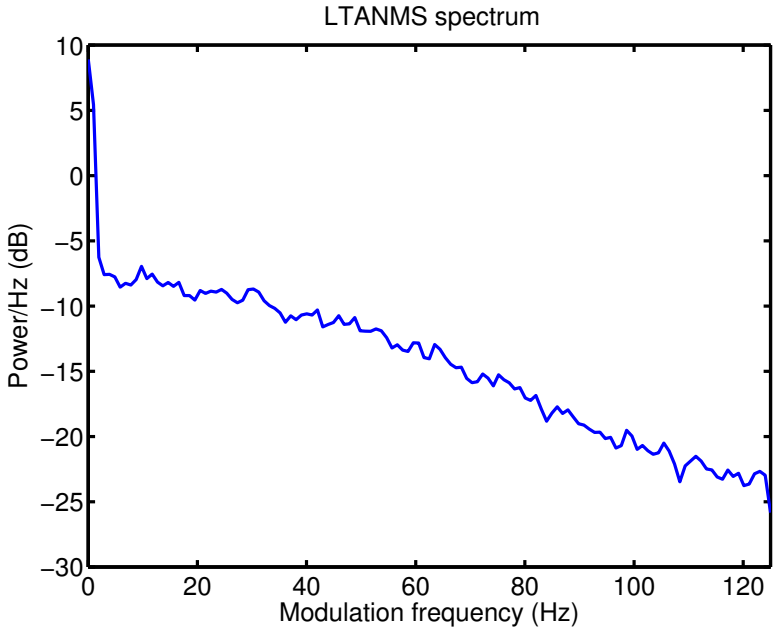

Figure 1.21.: LTANMS of street noise from RSG-10 noise database, which is obtained by averaging over about 65 seconds of street noise signal.

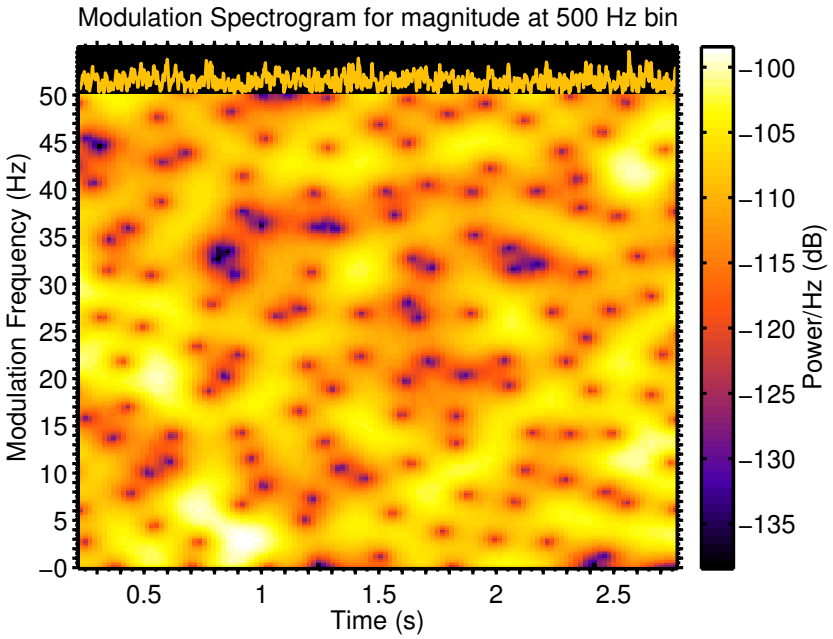

Figure 1.22.: Modulation spectrum of street noise from RSG-10 noise database.

Apart from the noises in the RSG-10 noise dataset and ITU-T test dataset, there is

another kind of noise which is referred to as speech-shaped noise [26]. The speech-shaped noise is a random noise that has the same long-term spectrum as a given speech signal, which is a stationary noise with colored characteristics. The spectrogram of the speech-shaped noise is shown in Figure 1.23. The LTANS, LTANMS and modulation spectrum are given in Figure 1.24, 1.25 and 1.26 respectively. Because speech-shaped noise is generated by filtering white noise with a filter whose spectrum equaled the long-term spectrum of the speech, its LTANS and LTANMS are similar as that of the speech. However in a short-time modulation frame, the speech-shaped noise has the similar characteristics as colored noise, thus its modulation spectrum in 1.26 is similar to that of noise.

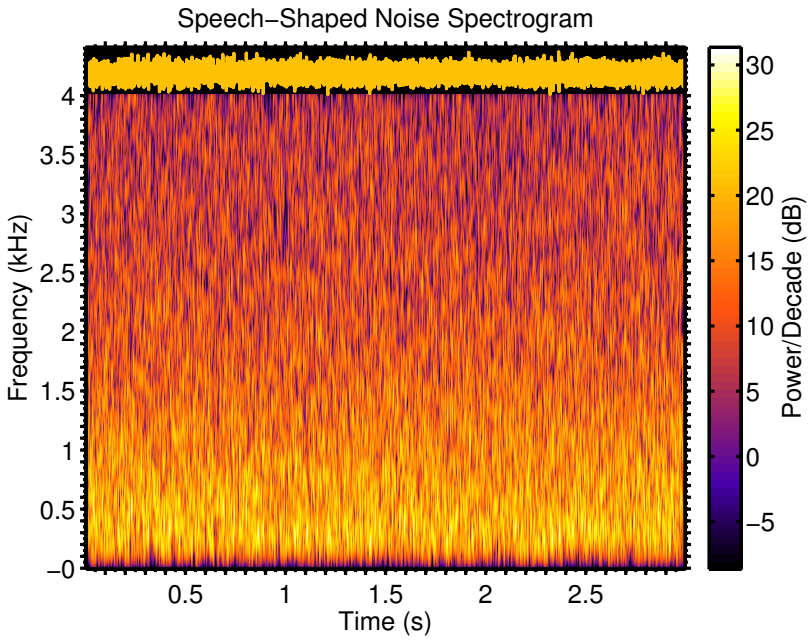

Figure 1.23.: Spectrogram of speech-shaped noise

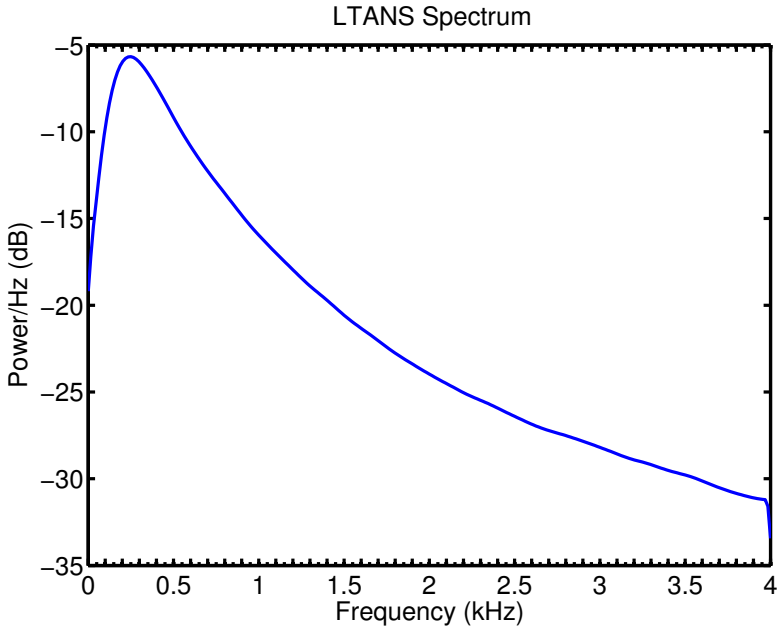

Figure 1.24.: LTANS of speech-shaped noise, which is obtained by averaging over about 65 seconds of speech-shaped noise signal.

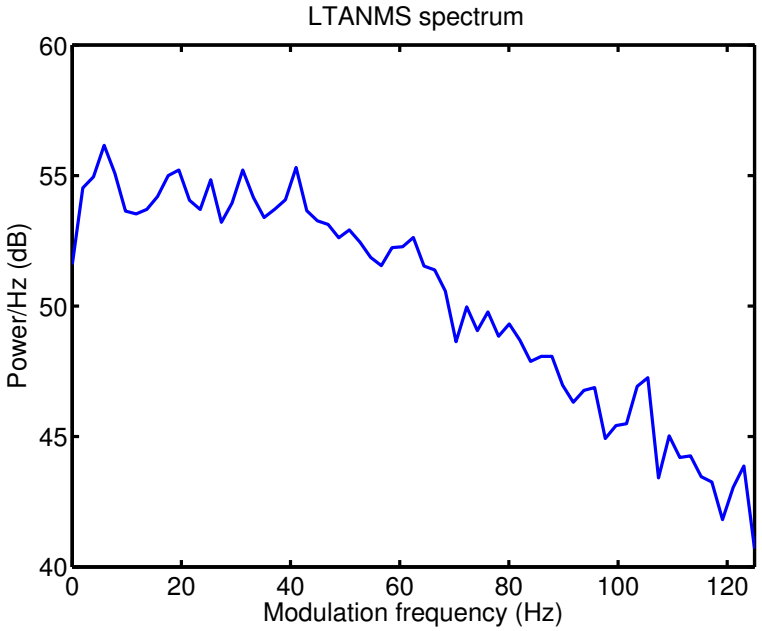

Figure 1.25.: LTANMS of speech-shaped noise, which is obtained by averaging over about 65 seconds of speech-shaped noise signal.

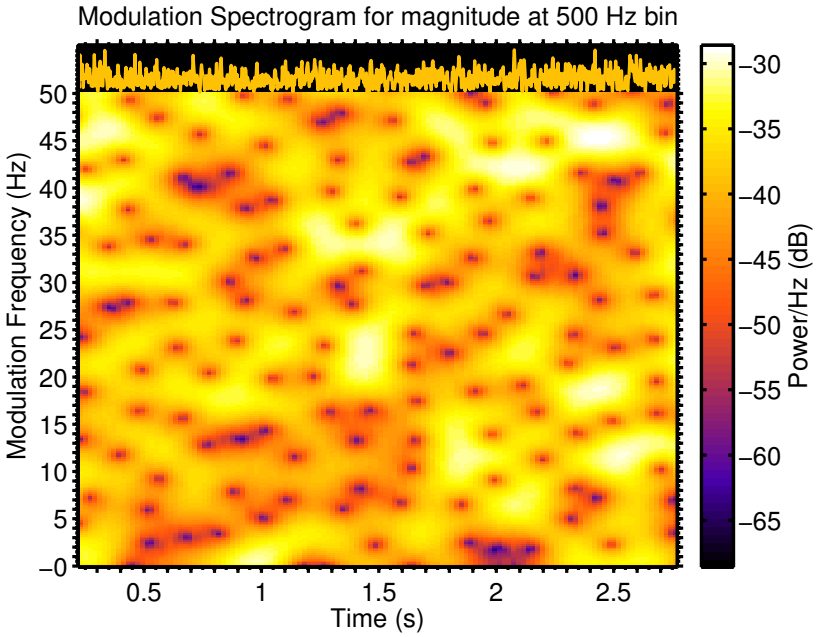

Figure 1.26.: Modulation spectrum of speech-shaped noise.

1.4.2.5. Modulation domain LPC of noise

The prediction gain of different LPC orders over acoustic frequencies are given from Figure 1.27 to Figure 1.29. The gains are calculated for white noise, car noise and street noise. The acoustic and modulation domain framing parameters are set in the same vein as the parameters for speech LPC model in Section 1.4.1.4. The length of the noises are 60 seconds and the acoustic frame increment is 4 ms, thus each kind of noise has 15000 acoustic frames involved in the averaging in (1.4). As can be seen from the figures, the LPC models with of order ≥ 3 orders are able to model the noises in the modulation domain. The prediction gains of white noise are about 15 dB over acoustic frequencies, which are fairly stable because of the stationary power distribution of white noise (the sudden drop of prediction gain at very low and very high frequencies results from the framing and windowing in the time domain). It worth nothing that the predictability of the spectral amplitudes

of the white noise results from the amplitudes correlation that is introduced by the overlapped windowing when doing the STFT. The derivation of the autocorrelation sequence of the spectral amplitude of white noise will be given in Section 4.3 in Chapter 4. For car noise, because nearly all of acoustic spectral power is distributed at low acoustic frequencies, the temporal acoustic sequences within these frequency bins are easier to predict from the previous acoustic frames, therefore the prediction gains are clearly higher at low frequencies than those at high frequencies, which are about 13 dB. For the street noise, the gains are fairly stable as was the case with white noise except at low frequencies (10 to 200 Hz), where the prediction gains are higher (over 15 dB) than those of higher frequencies. In the following chapters a modulation-domain LPC model of order 4 will be used when a noise LPC model is needed in the tested algorithms.

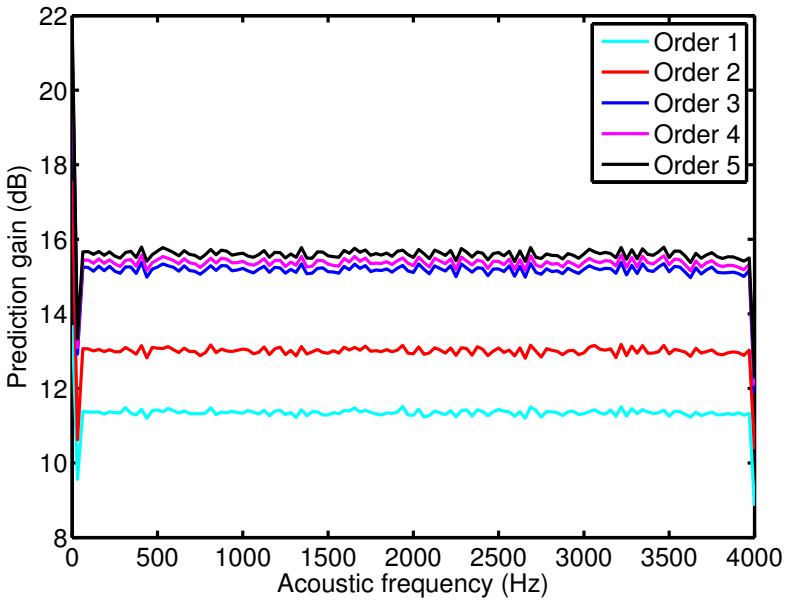

Figure 1.27.: Prediction gain of modulation-domain LPC model of different orders for white noise. The noise power and prediction error power are averaged over 15000 acoustic frames.

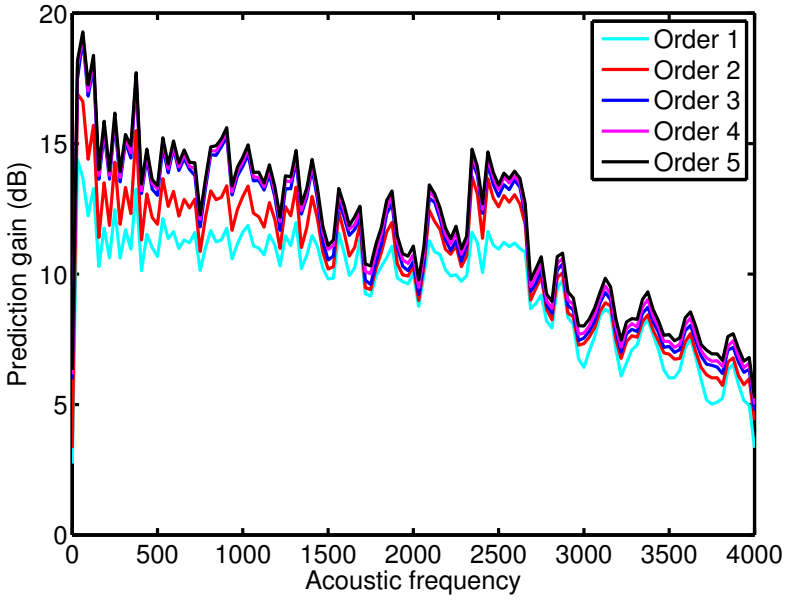

Figure 1.28.: Prediction gain of modulation-domain LPC model of different orders for car noise. The noise power and prediction error power are averaged over 15000 acoustic frames.

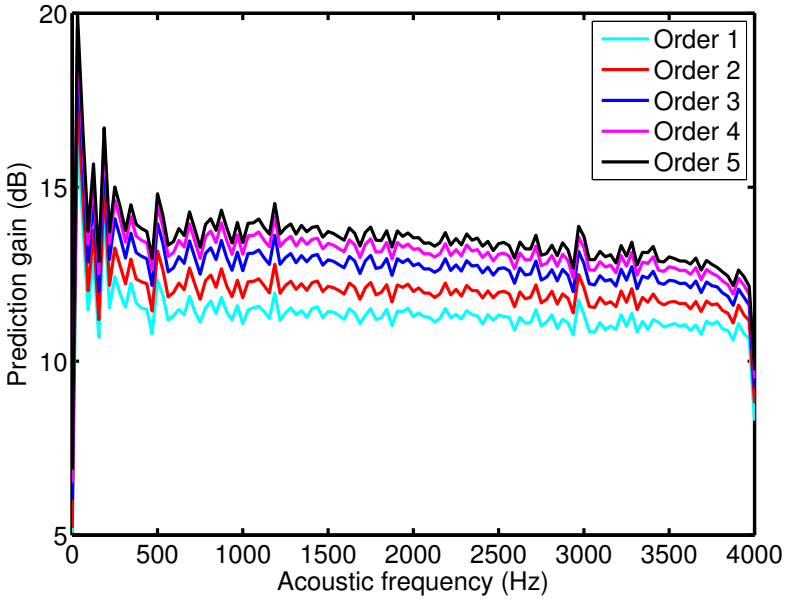

Figure 1.29.: Prediction gain of modulation-domain LPC model of different orders for street noise. The noise power and prediction error power are averaged over 15000 acoustic frames.

1.5. Thesis Structure

The focus in this thesis is on designing speech enhancement algorithms with better performance in improving speech quality, by incorporating the modulation domain characteristics into the time-frequency domain processing. The following chapters will describe the details of the algorithms which have been proposed:

Chapter 2 will give a literature review of speech enhancement algorithms including well-known and state-of-the-art algorithms. The types of the enhancer reviewed include time-frequency domain enhancers, subspace enhancers, modulation domain enhancers and post-processor-based enhancers. A number of relevant techniques, such as noise estimation and speech quality assessment, are also reviewed in this chapter.

Chapter 3 will describe a number of different post-processors using a modulation-domain Kalman filter. In the first part of the chapter, the modulation domain Kalman filter is introduced in the post-processing and two modified LPC models are also derived and incorporated into the Kalman filter. In the second part of the chapter, a Gaussian mixture noise model is incorporated in the Kalman filter which models the prediction error of the noise in the output spectral amplitude of a MMSE enhancer and based on this model.

Chapter 4 will present a speech enhancement algorithm using a subspace decomposition technique in the short-time modulation domain. In this algorithm, the modulation envelope of the noisy speech signal is decomposed into a signal space and a noise space. This decomposition is motivated by the predictability of the modulation envelope of the speech signal shown in Section 1.4.1.4.

Chapter 5 will propose two MMSE spectral amplitude estimators which incorporate the temporal dynamics of amplitude spectrum of speech and noise in the MMSE

estimation making use of a modulation domain Kalman filter. In the first part of the chapter, a MMSE spectral amplitude estimator assuming a generalised Gamma model for speech amplitude and Gaussian noise model is derived, the noise spectrum is pre-computed using a noise estimator. In order to incorporate the temporal dynamics of the noise amplitudes as well, in the second part of the chapter, a 'Gauss-ring' model is proposed under the assumption that the speech and noise amplitudes follow a Nakagami-m distribution and their phases are uniformly distributed..

Chapter 6 will summarise the thesis and give the possible ideas for extending the works.

2. Literature Review

2.1. Speech Enhancement

The objective of this chapter is to give an overview of the speech enhancement problem and the commonly used quality assessment methods. Speech enhancement is necessary in many applications, such as communication systems and hearing-aid devices. Over the past three decades, several classes of algorithm have been developed using a variety of mathematical models and techniques. Comprehensive overviews of speech enhancement are available in review papers [27, 3] and textbooks [1, 28]. Speech degradations may involve a combination of additive noise, convolutive effects and non-linear distortion. The research in this thesis focuses on solving the single-channel speech enhancement problem where speech signals are corrupted by additive noise. The aim is to solve the problem when only one signal channel is available.

2.2. Noise Power Spectrum Estimation

Noise estimation plays a important role in speech enhancement algorithms and the performance of most enhancers is significantly affected by the noise estimation technique. For many speech enhancement methods, a necessary first step is to estimate

the time-averaged power spectrum of the interfering additive noise. The estimation of the noise is often performed in the spectral domain because 1) the spectral components of speech and noise can be partially decorrelated; and 2) Because most of the spectral power of speech and most of the types of noise lies within specific frequency bands, the speech and noise are often sparse in the spectral domain which makes them easier to separate. In order to distinguish between the speech and noise components of the single input channel, it is necessary to use prior information about how they differ. The most common assumptions are the speech has higher energy than the noise and/or that the speech is less stationary than the noise. There are several classes of noise estimation algorithm based on different techniques, which will be reviewed in the following.

2.2.1. Voice activity detection

A straightforward way to estimate the noise spectrum is to use a Voice Activity Detector (VAD) to identify when speech is absent and to update the noise estimate during these periods. The update of the noise often relies on a smoothing constant and is given by

$$|\widehat{W}_{n,k}|^2 = \kappa_{n,k}|\widehat{W}_{n-1,k}|^2 + (1-\kappa_{n,k})|Z_{n-1,k}|^2 \tag{2.1}$$

where $|\widehat{W}_{n,k}|^2$ is the estimate of the noise power spectrum. $\kappa_{n,k}$ is the smoothing constant, which is named forgetting factor. The smoothing operation is applied in order to reduce the variance in the estimated noise power spectrum. The value of the forgetting factor, which is normally in the range 0.5 to 0.9, determines the number of frames involved in the averaging of the noise estimate. For instance, when $\kappa = 0.9$, the noise estimate is averaged over 20 acoustic frames [29]. Therefore, the forgetting factor can be used to control the trade-off between the tracking capability and the

variance of the noise estimation. For noise that is non-stationary across time and frequencies, $\kappa_{n,k}$ is normally selected differently for each time-frequency cell, as will be explained later in this section.

A VAD normally operates by extracting features (e.g. energy levels, pitch, zero crossing rate and cepstral features) from the noisy speech and, based on these, determining the speech absence using specified decision rules. The performance of the VAD depends on the SNR of the noisy speech and when the SNR is very low and the noise is non-stationary its performance is normally degraded [30]. Many VADs have been proposed for speech enhancement based on a range of features, models and decision rules. The short-time signal energy is one of the earliest features in the design of VADs [31]. The frame energies of the noisy speech signal are calculated and compared with a threshold, under the assumption that the energy of the frames where speech is present will be significantly larger than those where there is no speech. The threshold can either be predetermined or else chosen adaptively. In [32], the threshold is selected to be at the 80th centile of the histogram of the energies that are below an upper preset maximum threshold. In addition to the short-time energy, two other features which were often used in a VAD are zero crossing rate and period. The zero-crossing rate is the number of times the successive samples in a speech signal passes through the value of zero; this is effective at identifying noise that has significant energy at high frequencies but can also falsely identify some speech sounds as noise. The VAD defined in the G.729 standard [33] is widely used in speech processing applications, and is based on four features: low-band and full-band energy, line spectral pairs and zero-crossing rate. In the G.729 codec, an initial VAD is firstly obtained which is then smoothed according to the stationary nature of the speech and interference. Additionally, periodicity of the signals can also be used in the design of a VAD because unlike speech signal, most of noise

signals are aperiodic. The main difficulty in using periodicity is that the VAD does not work for periodic noises [34].

Rather than the feature-based VAD methods, it is also possible to design VADs by modelling the transformed coefficients of the signals. The VAD presented in [30] employs Gaussian distributions to model the complex STFT coefficients of the speech, noise and clean speech and the decision rule is based on likelihood ratio test. The parameters in the models are estimated using a decision directed method. It is shown in [30] that this method performs consistently better than G.729 for different kinds of noise at low SNRs for speech frame detection .

Instead of making a hard decision for VAD, there is also a class of methods which applies a soft-decision VAD using a forgetting factor $\kappa_{n,k}$ that varies according to the Speech Presence Probability (SPP). The reason for applying a SPP is that speech may not be present in every spectral component of a frame. The SPP can be estimated for different time frames based on the features such as averaged SNR over all frequencies [35] and the ratio between the local energy of the noisy speech and its minimum within a specified time frame [36]. The method in [35] proposes a frequency-dependent factor which depends on the estimated speech presence combined with the estimated SNR averaged over a short time to control the forgetting factor. In [36], a speech probability is estimated depending on the ratio between the power in the current frame and its minimum within a specified frame. This approach is extended in [37] which suggests a two-step procedure that modifies an initial speech presence estimation. This method is further extended in [38, 39] which have a lower latency and a frequency-dependent threshold on the ratio of noisy speech power to minimum power in order to estimate the speech presence probability.

2.2.2. Minimum statistics

Since voice activity detection is difficult in non-stationary noise scenarios, especially at low SNR scenarios, there are other noise estimation approaches which do not make use of a VAD. One representative method is the Minimum Statistics (MS) method proposed in [40] and modified in [41]. The assumption in this method is that, in any given frequency bin, there will be times when there is little speech power and that the power of the noisy signal will then be dominated by the noise. The noise power can therefore be estimated by tracking the minimum power within a past time period (typically 0.5 to 1.5 seconds). As discussed in Section 2.2.2.1, because the output of the minimum filter underestimates the true noise power, a bias compensation factor is needed and in [41], an approximation of the compensation factor which varies with time and frequency is proposed. A more complete analysis of the factors that contribute to the bias of the MS estimate is given in [42] and a number of efficient approximations are proposed therein.

2.2.2.1. MMSE estimation

The main drawback of the MS algorithm is that when the noise power increases during the interval over which the minimum is taken, it will be underestimated or tracked with some delay [43]. More recently, a number of MMSE-based noise estimation algorithms have been proposed [44, 43]. In [44], an MMSE noise estimator and an associated bias compensation factor are derived under the assumption that the complex STFT coefficients follow a complex-Gaussian distribution. The bias factor is derived as a function of the a priori SNR, which is estimated using the direct-decision method [7]. Compared to the MS algorithm [41], this method not only has a lower computational complexity, but also gives better performance in noise tracking and speech enhancement [44]. A more recent MMSE-based approach

extending this algorithm is proposed in [43]. In this work, it is shown that the noise estimator in [44] can actually be interpreted as a hard VAD-based estimator where the VAD is determined by the comparison between the spectral power of the noisy speech at current frame and that of the noise at previous frame. The estimator in [43] improves the performance of [44] by replacing the VAD by a soft speech presence probability with fixed priors. Additionally, because this estimator does not need the evaluation of a biased factor, it is more computationally efficient than [44] where an incomplete Gamma function needs to be evaluated to determine the bias compensation factor.

2.3. Subspace Enhancement

The subspace method of speech enhancement was firstly proposed in [23]. Its key assumption is that speech is generated by a low-order autogressive model and that the samples in a frame of speech therefore lie within a low-dimensional subspace. The space of T-dimensional noisy speech vectors can be decomposed into a M-dimensional $(M < T)$ *signal subspace* containing both speech and noise and a $(T-M)-$dimensional *noise subspace* containing only noise; the aim of the subspace enhancement is to identify this subspace and constrain the clean speech samples to lie within it. If the noise is white, the decomposition can be obtained by applying the Karhunen-Loéve Transform (KLT) [45] to the noisy speech covariance matrix. KLT components represent the variance along each of the principle components, which are the eigen-vectors of the covariance matrix of the signal. After the KLT components representing the signal subspace and noise subspace are obtained, the KLT components representing the signal subspace are modified by a gain function determined by the estimator. The linear estimator minimizes the speech signal dis-

tortion while applying either a Time Domain Constraint (TDC) or Spectral Domain Constraint (SDC) to the residual noise energy. The TDC and SDC criterion differ in the domains where the constraint is applied when making the optimal estimation [23]. If T-dimensional vector of speech and noise signal of frame n are defined as $\mathbf{s}_n$ and $\mathbf{w}_n$ respectively, and assume that the estimator for the frame is a $T \times T$ matrix, $\mathbf{H}_n$, thus the estimate of the clean speech vector is obtained as

$$\hat{\mathbf{s}}_n = \mathbf{H}_n \left(\mathbf{s}_n + \mathbf{w}_n\right) = \mathbf{H}_n \mathbf{z}_n$$

The residual signal is defined as the difference between the clean speech and its estimation

$$\begin{aligned} \mathbf{r}_n &= \hat{\mathbf{s}}_n - \mathbf{s}_n \\ &= \left(\mathbf{H}_n - \mathbf{I}\right) \mathbf{s}_n + \mathbf{H}_n \mathbf{w}_n \\ &\triangleq \mathbf{r}_s + \mathbf{r}_w \end{aligned}$$

where $\mathbf{r}_s$ represents signal distortion and $\mathbf{r}_w$ represents the residual noise. The optimal estimation is derived by minimizing the signal distortion energy of the frame, ϵ_s^2 and for the TDC, this subjects to the constraint that the residual noise energy, ϵ_w^2, is smaller than a preset value, which are defined as

$$\min_{\mathbf{H}} \epsilon_s^2 \qquad \text{subject to } \frac{1}{N}\epsilon_w^2 \leq \alpha \sigma_w^2 \tag{2.2}$$

where N is the length of the speech signal frame, σ_w^2 is the noise variance and $0 \leq \alpha \leq 1$ is a constant controlling amount of the residual noise. The solution to this optimization problem, known as the TDC estimator, is given by [23]

$$\mathbf{H}_{\text{TDC}} = \mathbf{R}_S \left(\mathbf{R}_S + \eta \mathbf{R}_W\right)^{-1} \tag{2.3}$$

where $\mathbf{R}_S$ is the covariance matrix of the clean speech and $\mathbf{R}_W$ is the covariance matrix the noise. η is the Lagrange multiplier, which satisfies

$$\alpha = \frac{1}{N} \text{tr} \left\{ \mathbf{R}_S^2 \left(\mathbf{R}_S + \eta \sigma_w^2 \mathbf{I} \right)^{-2} \right\}. \tag{2.4}$$

For white noise, the estimator in (2.3) becomes

$$\mathbf{H}_{\text{TDC}} = \mathbf{R}_S \left(\mathbf{R}_S + \eta \sigma_w^2 \mathbf{I} \right)^{-1}$$

and applying the eigen-decomposition $\mathbf{R}_S = \mathbf{U} \Lambda \mathbf{U}^T$, where $\mathbf{U}$ represents the matrix of eigenvectors and Λ is a diagonal matrix with elements being the corresponding eigenvalues λ_i where $\lambda_i > 0$ for $i = 1 \cdots M$ and $\lambda_i > 0$ for $i = M + 1 \cdots T$. The estimator now becomes

$$\mathbf{H}_{\text{TDC}} = \mathbf{U} \Lambda \left(\Lambda + \eta \sigma_w^2 \mathbf{I} \right)^{-1} \mathbf{U}^T \tag{2.5}$$

As can been seen from the estimator in (2.5), η can determine the residual noise and signal distortion more intuitively than α. As a result, η is normally specified instead of α. In order to compromise between residual noise and signal distortion, η can be set according to the SNR [23, 46]. Good estimate of $\mathbf{R}_Z$ and $\mathbf{R}_W$ is important for calculating the estimator in (2.5). The estimate of $\mathbf{R}_Z$ can be obtained from the empirical covariance of non-overlapping vectors of the noisy speech signal in the neighborhood of the current sample, $\mathbf{z}_n$. $\mathbf{R}_W$ is often estimated from vectors of the noisy speech signal during which speech is absent [23, 46]. In Section 4.3, a new

method to estimate $\mathbf{R}_W$ in the modulation domain will be proposed.

The above equations are derived for white noise. If the speech is degraded by colored noise, it can be firstly whitened using a linear transform $\mathbf{R}_W^{-\frac{1}{2}}$ based on an estimate of $\mathbf{R}_W$. In this case, the TDC estimator in (2.5) becomes

$$\mathbf{H}_{\text{TDC}} = \mathbf{R}_W^{\frac{1}{2}} \mathbf{U} \Lambda \left(\Lambda + \eta \mathbf{I}\right)^{-1} \mathbf{U}^T \mathbf{R}_W^{-\frac{1}{2}} \tag{2.6}$$

For the SDC estimator, the constraint in (2.2) is applied to the spectrum of the residual noise. The estimator is proposed as

$$\begin{aligned} &\min_{\mathbf{H}} \epsilon_s^2 \\ \text{subject to } &\mathrm{E}\left(\mathbf{u}_i^T \mathbf{r}_w\right) \leq \alpha_i \sigma_w^2 \quad i = 1, \ldots, M \\ \text{and } &\mathrm{E}\left(\mathbf{u}_i^T \mathbf{r}_w\right) = 0 \quad i = M+1, \ldots T \end{aligned}$$

where $\mathbf{r}_w$ is the time-domain residual noise vector and $\mathbf{u}_i$ is the ith column of the eigen-vector matrix $\mathbf{U}$, thus $\mathbf{u}_i^T \mathbf{r}_w$ is the ith KLT components of $\mathbf{r}_w$. The solution of the SDC estimator, $\mathbf{H}_{\text{SDC}}$, is given by

$$\begin{aligned} \mathbf{H}_{\text{SDC}} &= \mathbf{U}\mathbf{V}\mathbf{U}^T \\ \mathbf{V} &= \text{diag}\left(v_{11} \ldots v_{TT}\right) \\ v_{ii} &= \begin{cases} \sqrt{\alpha_i} & i = 1, \ldots, M \\ 0 & i = M+1, \ldots, T \end{cases} \end{aligned}$$

α_i can be selected independently of the statistics of the speech and noise. Two possible choices for α_i are given by [23]

$$\alpha_i = \left(\tfrac{\lambda_i}{\lambda_i + \sigma_w^2}\right)^g \text{and } \alpha_i = \exp\left\{\tfrac{-c\sigma_w^2}{\lambda_i}\right\}$$

where $g \geq 1$ and $c \geq 1$ are experimentally determined constant.

Although the estimator for the white noise shown above can be used to remove colored noise making the use of pre-whitening, the covariance matrix of some noises, such as narrowband noise, is rank deficient. To solve this problem, an approach is proposed in [47]. In this approach, the noisy speech frames are classified into speech dominated frames and noise dominated frames. For the noise dominated frames, the eigenvectors of the noise covariance matrix and those of the speech can be assumed to be identical because speech spectrum is flatter in these frames. This approach does not require noise whitening and provides better noise shaping. In a generalization of the method, [46] applies a non-unitary transformation to the noisy speech vectors that simultaneously diagonalizes the covariance matrices of both speech and colored noise. However, unlike the algorithm in [48], it does not give an explicit solution to the SDC estimator. The SDC estimator in [48] extends the subspace algorithm in [49] to colored noise and derives the explicit solution for the SDC estimator.

2.4. Enhancement in the Time-Frequency Domain

Two influential enhancers in the time-frequency domain are the spectral subtraction method in [50] and MMSE spectral amplitude estimator in [7]. The spectral subtraction method is still one of the most popular methods of noise reduction, where the estimated magnitude or power spectrum of the noise is subtracted from that of the noisy speech. The general gain function in the STFT domain is given as

$$G_{ss}(n,k) = \max\left\{\frac{(|Z_{n,k}|^r - |\hat{W}_{n,k}|^r)^{1/r}}{|Z_{n,k}|}, 0\right\} \tag{2.7}$$

where $|\hat{W}_{n,k}|$ is the estimated noise amplitude spectrum and r determines the domain of the subtraction operates. It has been found that when $r = 2$ the method performs

the best [51]. However, $r = 1$ is more commonly used and gives more noise reduction at poor SNRs. Although this algorithm can reduce the unwanted noise dramatically, residual broadband noise and musical noise remain in the enhanced speech. To improve the performance of the spectral subtraction method, a spectral floor and oversubtraction is introduced, which leads to a modified STFT-domain gain

$$G_{ss}(n,k) = \max\left\{\frac{(|Z_{n,k}|^r - \alpha|\hat{W}_{n,k}|^r)^{1/r}}{|Z_{n,k}|}, \psi|\widehat{W}_{n,k}|\right\} \tag{2.8}$$

where $\alpha \geq 1$ and $0 \leq \psi \ll 1$ are factors controlling the oversubtraction and the noise floor respectively. The oversubtraction leads to an attenuation of the residual noise by reducing the spectral excursions in the speech spectrum, but it may introduce distortion of the speech if it is set too high. The parameter ψ controls the spectral floor of the enhanced speech, which retains a small amount of the original noisy signal to reduce the perception of the musical noise. This is because, as mentioned is Chapter 1, in the time-frequency domain, musical noise exists as isolated spectral peaks; applying a spectral floor can fill the valley between the large peaks and thus reduce the apparent musical noise. Because the noise power is often assumed to be constant and the speech power is non-stationary in different frames, α is often varied in each frame according to the SNR in each frame so that less subtraction is performed in frames with high SNR [52]. The algorithm in [29] extends this method, it controls both the oversubtraction and the noise floor adaptively based on a perceptual threshold function, which is more closely correlated with speech perception than the SNR.

Another influential algorithm in the time-frequency domain is the MMSE spectral amplitude estimator in [7]. In this algorithm, the assumptions about the speech and noise models in the time-frequency domain are:

1. The complex STFT coefficients of speech and noise are additive,

2. The spectral amplitudes of speech follow a Rayleigh distribution (prior distribution),

3. The additive noise is Gaussian distributed (observation distribution).

Under these assumptions, the posterior distribution of the spectral amplitudes of speech has a Rician distribution. The estimator can be derived by minimizing the mean-square error between the estimated amplitude and the clean speech amplitude, which is given by the mean of the posterior distribution. The gain function of each time-frequency bin is given by [7]

$$\begin{aligned} G_{mmse}(n,k) &= \Gamma(1.5)\frac{\sqrt{\upsilon_{n,k}}}{\zeta_{n,k}}\mathcal{M}(-0.5;1;-\upsilon_{n,k}) \\ &= \Gamma(1.5)\frac{\sqrt{\upsilon_{n,k}}}{\zeta_{n,k}}\exp\left(-\frac{\upsilon_{n,k}}{2}\right)\left[(1+\upsilon_{n,k})I_0\left(\frac{\upsilon_{n,k}}{2}\right)+\upsilon_{n,k}I_1\left(\frac{\upsilon_{n,k}}{2}\right)\right] \end{aligned} \tag{2.9}$$

where $\Gamma(\cdot)$ is the gamma function and $\mathcal{M}$ is the confluent hypergeometric function (see Appendix A). I_0 and I_1 denote the modified Bessel function of zero and first order, respectively. $\upsilon_{n,k}$ is defined as

$$\upsilon_{n,k} = \frac{\xi_{n,k}}{1+\xi_{n,k}}\zeta_{n,k} \tag{2.10}$$

where $\xi_{n,k}$ is interpreted as the a priori SNR, which is defined as the ratio of the variances of the kth spectral component of the speech to that of the noise, $\nu^2(n,k)$ while $\zeta_{n,k}$ is referred to as the a posteriori SNR which is the ratio $\frac{R^2(n,k)}{\nu^2(n,k)}$. As can be seen, central to calculation of the gain in (2.9) is the estimation of the a priori SNR and in [7] a "decision directed" approach is proposed, where $\xi_{n,k}$ is estimated as

$$\hat{\xi}_{n,k} = \tau\frac{\hat{A}^2(n-1,k)}{|\hat{W}(n-1,k)|^2} + (1-\tau)\max\left(\zeta_{n,k}-1,0\right), \quad 0 \leq \tau < 1 \tag{2.11}$$

where $\hat{A}(n,k)$ is the estimated amplitude of the kth signal spectral component in the nth frame and τ is a temporal smoothing constant. The MMSE enhancer in [7] is improved in [53] by using the mean-square error of the estimated log amplitude as the distortion measure, and it has been found that this gives slightly improved speech quality. Assuming the same statistical models as those in [7], the gain function of the logMMSE estimator is derived as

$$G_{\text{logmmse}}(n,k) = \frac{\xi_{n,k}}{1+\xi_{n,k}} \exp\left\{\frac{1}{2}\int_{\upsilon_{n,k}}^{\infty} \frac{e^{-t}}{t} dt\right\} \tag{2.12}$$

It is claimed that this estimator can give low background noise levels without introducing additional distortion.

The drawback of the MMSE enhancer is that when the correlation length of the speech signal is longer than the frame length, the spectral coefficients of the speech do not follow a Gaussian distribution and the spectral outliers will therefore introduce artefacts. A number of papers, based on different statistical models, are proposed to model the spectral amplitude or complex-valued coefficients. The papers [54, 55] derived estimators based on the MMSE and Maximum a Posteriori (MAP) criterion, respectively. The main contribution of [54] is that, instead of using Gaussian Probability Density Function (PDF), it introduces super-gaussian distributions (complex Laplacian and Gamma PDF) to model the PDF of the real and imaginary parts of the complex STFT coefficients of speech and complex Gaussian and Laplacian PDF for the coefficients of the noise. It is found that the estimators based on the supergaussian models outperform the amplitude-domain MMSE estimators [7] since they give higher SNR improvements. However, when both speech and noise are modeled by supergaussian PDFs, there is no exact analytic solution given in [54] for amplitude estimation. To solve this problem, a computationally simpler MAP magnitude estimator is derived in [55] which approximates the MMSE estimator in

this case. It is found that the introduction of the supergaussian models can result in less musical noise in the estimated speech. In the same vein, a three-parameter generalized Gamma prior is assumed in [56] when estimating the STFT magnitude and complex-valued STFT coefficients, which is given by

$$p(a) = \frac{d a^{d\gamma-1}}{\beta^{2\gamma}\Gamma(\gamma)} \exp\left(-\frac{a^d}{\beta^2}\right) \tag{2.13}$$

where $\gamma > 0$, $\beta > 0$ and $d > 0$ are the three parameters. The distribution in (2.13) includes some special cases, for example, when $\gamma = 1$ it becomes the Weibull distribution and when $d = 2$ and $\gamma = 1$, it becomes the Rayleigh distribution. Therefore, the complex STFT estimators and spectral amplitude estimators derived using (2.13) in [56] can be seen as a generalized case that includes the estimators in [7] and [54] as special cases. In [56], the two cases $d = 1$ and $d = 2$ are exploited in deriving the MMSE estimators for the amplitudes and the real and imaginary parts of the speech STFT coefficients and when estimating the complex STFT coefficients, a two-sided version of the distribution in (2.13) is considered. For $d = 1$, a closed form cannot be obtained and thus two approximations are proposed in [56] under different SNR conditions while for $d = 2$ a closed form is derived. It is shown that the amplitude estimators derived using the distributions in (2.13) are slightly better than MAP estimator in [57] in that the speech distortion that is introduced is slightly less. In Chapter 5, the distribution when $d = 2$ will be used in deriving an MMSE estimator.

Rather than assuming different statistical models for the speech and noise, some methods have been proposed which modify the cost function used in the derivation of the estimators. The mean squared-error cost function used in [7, 53] is not perceptually meaningful in that it does not necessarily produce estimators that emphasize spectral peak information or estimators which take into account auditory

masking effects [58]. Therefore, the cost functions proposed are normally designed to reflect the perceptual characteristics of speech and noise. For example, in [59, 8], masking thresholds are incorporated into the derivation of the optimal spectral amplitude estimators. The threshold for each time-frequency bin is computed from a suppression rule based on an estimate of the clean speech signal. It is shown that this estimator outperforms the MMSE estimator [7] with less musical noise. On the other hand, in [58] different distortion measures are used in the cost function, which include four types of measures: weighted Euclidean (WE) distortion measure, Itakura-Saito (IS) measure, COSH measure [60] and Weighted Likelihood Ratio (WLR) measure. The definition of the three measures are given by:

$$
\begin{aligned}
d_{\mathrm{WE}}\left(|S_{n,k}|,|\widehat{S}_{n,k}|\right) &= |S_{n,k}|^{u}\left(|S_{n,k}|-|\widehat{S}_{n,k}|\right)^{2} \\
d_{\mathrm{IS}}\left(|S_{n,k}|^{2},|\widehat{S}_{n,k}|^{2}\right) &= \frac{|S_{n,k}|^{2}}{|\widehat{S}_{n,k}|^{2}}-\log\left(\frac{|S_{n,k}|^{2}}{|\widehat{S}_{n,k}|^{2}}\right)-1 \\
d_{\mathrm{COSH}}\left(|S_{n,k}|,|\widehat{S}_{n,k}|\right) &= \frac{1}{2}\left[\frac{|S_{n,k}|}{|\widehat{S}_{n,k}|}+\frac{|\widehat{S}_{n,k}|}{|S_{n,k}|}\right] \\
d_{\mathrm{WLR}}\left(|S_{n,k}|,|\widehat{S}_{n,k}|\right) &= \left(\log|S_{n,k}|-\log|\widehat{S}_{n,k}|\right)\left(|S_{n,k}|-|\widehat{S}_{n,k}|\right)
\end{aligned}
\tag{2.14}
$$

where u is a power exponent. When $u>0$, the distortion measure d_{WE} emphasizes spectral peaks, while when $u<0$, this distortion measure emphasizes spectral valleys. It is found that the amplitude estimators that emphasize spectral valleys more than the spectral peaks performed the best in terms of having less residual noise and better speech quality among all the estimators. When $u=-1$, the resultant estimators outperform the MMSE estimator with a 70% preference in a subjective listening test. A generalized cost function is proposed in [61], based on which a β-order MMSE estimator is derived. Here β represents the order of the spectral amplitude used in the calculation of the cost function. In this work, the relation between β and the spectral gain function is firstly investigated. Also, they propose

an adaption method for β, which is calculated according to the SNR of the frame. The performance of this estimator is shown to be better than both the MMSE estimator and the logMMSE estimator in that it gives better noise reduction and better estimation of weak speech spectral components. The estimators in [58] and [61] are extended in [62], where a weighted β-order MMSE estimator is proposed. It employs a cost function which combines the β-order compression rule and the WE cost function. The parameters β and u are selected based on the characteristics of the human auditory system. It is shown that the modified cost function leads to a better estimator giving consistently better performance in both subjective and objective experiments, especially for noise having strong high-frequency components and at low SNRs.

2.5. Enhancement in the Modulation Domain

There is increasing evidence, both physiological and psychoacoustic, to support the significance of the modulation domain in speech enhancement. Drullman et al. conducted experiments to study the intelligibility of speech signals with temporally modified spectral envelopes by applying low-pass and high-pass filters and they found that modulation frequencies between 4 Hz and 16 Hz have high contributions to the intelligibility of speech [18, 19], and that there is no significant linguistic information in either the very slow or the very fast components of the spectral envelopes of speech. Based on this observation, Hermansky et al. proposed a relative spectral (RASTA) technique which suppresses the fast and slow spectral components of speech signal by employing band-pass filtering of time trajectories of each frequency channel [63]. In 2007, Singh and Rao extended the technique which combined the framework of spectral subtraction with RASTA filtering and it is stated that this

approach can outperform both the spectral subtraction method and the RASTA speech enhancement method [64]. Additionally, Paliwal et al. recently proposed a series of speech enhancement algorithms which extended time-frequency domain algorithms to the modulation domain based on STFT analysis [65, 66, 67]. These methods involve spectral subtraction, Kalman filtering and MMSE estimation. They claim in the papers that these methods can outperform the original enhancers that apply the corresponding methods in the time-frequency domain and there is less musical noise in the speech enhanced by these methods. The modulation domain is also important in the area of speech intelligibility metrics where Taal et al. [17] proposed recently a STOI metric. This metric calculates the sample correlation coefficient between the short-time temporal envelope of the clean speech and that of the noisy speech as an intermediate intelligibility measure. This intelligibility measure shows high correlation with the intelligibility of time-frequency weighted noisy speech which outperforms four widely used measures in predicting the intelligibility.

In the rest of the section the modulation domain Kalman filter and the modulation domain spectral subtraction will be introduced in detail because it will be used in subsequent chapters.

2.5.1. Modulation domain Kalman filtering

In [66], the author assumes an additive model of the noisy speech amplitude, which is

$$|Z_{n,k}| = |S_{n,k}| + |W_{n,k}| \tag{2.15}$$

where n denotes the acoustic frame and k denotes the acoustic frequency bin. To perform Kalman filtering in the modulation domain, each frequency bin is processed independently; for clarity, the frequency index, k, will be omitted in the description

that follows.

Assuming that the temporal envelope, $|S_n|$, of the amplitude spectrum of the speech signal can be modeled by a linear predictor with coefficients $\mathbf{b} = [b_1\, b_2 \cdots b_p]$ in each modulation frame:

$$|S_n| = -\sum_{i=1}^{p} b_i |S_{n-i}| + \tilde{e}_n \tag{2.16}$$

where $\tilde{e}_n$ is assumed to be a random Gaussian excitation signal with variance $\tilde{\sigma}^2$. Since any type of noise is colored in the modulation domain because of the overlap between the acoustic frames, in [66] a Kalman filter for removing a colored noise is used [68]. The state vector of speech is augmented with the state vector of noise, and both of the speech and noise components are estimated simultaneously.

Within each frequency bin, the authors use autoregressive models for the speech and the noise of orders p and q respectively and so the state vector in the Kalman filter has dimension $p + q$. The dynamic model of the state space is given by

$$\begin{bmatrix} \tilde{\mathbf{s}}_n \\ \breve{\mathbf{s}}_n \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{A}}_n & \mathbf{0} \\ \mathbf{0} & \breve{\mathbf{A}}_n \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{s}}_{n-1} \\ \breve{\mathbf{s}}_{n-1} \end{bmatrix} + \begin{bmatrix} \tilde{\mathbf{d}} & \mathbf{0} \\ \mathbf{0} & \breve{\mathbf{d}} \end{bmatrix} \begin{bmatrix} \tilde{e}_n \\ \breve{e}_n \end{bmatrix}, \tag{2.17}$$

where $\tilde{\mathbf{s}}_n = [|S_n| \cdots |S_{n-p+1}|]^T$ is the speech state vector, $\tilde{\mathbf{d}} = [1\, 0 \cdots 0]^T$ is a p-dimensional vector and the speech transition matrix has the form $\widetilde{\mathbf{A}}(n) = \begin{bmatrix} -\tilde{\mathbf{b}}^T \\ \mathbf{I} \quad \mathbf{0} \end{bmatrix}$ where $\tilde{\mathbf{b}} = [b_1 \cdots b_p]^T$ is the LPC coefficient vector, $\mathbf{I}$ is an identity matrix of size $(p-1) \times (p-1)$ and $\mathbf{0}$ denotes an all-zero column vector of length $p-1$. The quantities $\breve{\mathbf{d}}$, $\breve{\mathbf{s}}_n$ and $\breve{\mathbf{A}}_n$ are defined similarly for the order-q noise model. (2.17) is re-written as

$$\mathbf{s}_n = \mathbf{A}_n \mathbf{s}_{n-1} + \mathbf{D}_1 \mathbf{e}_n, \tag{2.18}$$

where $\mathbf{s}_n$, $\mathbf{A}_n$ and $\mathbf{D}_1$ represent the composite speech+noise elements of as (2.17). The observation model is

$$|Z_n| = \begin{bmatrix} \tilde{\mathbf{d}}^T & \breve{\mathbf{d}}^T \end{bmatrix} \mathbf{s}_n = \mathbf{D}_2 \mathbf{s}_n \tag{2.19}$$

The equations of modulation domain Kalman filter are given by

$$\begin{aligned}
\boldsymbol{\Sigma}_{n|n-1} &= \mathbf{A}_n \boldsymbol{\Sigma}_{n-1|n-1} \mathbf{A}_n^T + \mathbf{D}_1 \mathbf{Q}_{n-1} \mathbf{D}_1^T && (2.20)\\
\mathbf{k}_n &= \boldsymbol{\Sigma}_{n|n-1} \mathbf{D}_2 [\mathbf{D}_2^T \boldsymbol{\Sigma}_{n|n-1} \mathbf{D}_2]^{-1} && (2.21)\\
\mathbf{s}_{n|n-1} &= \mathbf{A}_n \mathbf{s}_{n-1|n-1} && (2.22)\\
\boldsymbol{\Sigma}_{n|n} &= \boldsymbol{\Sigma}_{n|n-1} - \mathbf{k}_n \mathbf{D}_2^T \boldsymbol{\Sigma}_{n|n-1} && (2.23)\\
\mathbf{s}_{n|n} &= \mathbf{s}_{n|n-1} + \mathbf{k}_n [|Z_n| - \mathbf{D}_2 \mathbf{s}_{n|n-1}] && (2.24)
\end{aligned}$$

where $\boldsymbol{\Sigma}_{n|n}$ is the covariance matrix corresponding to the estimates, $\mathbf{Q} = \begin{bmatrix} \tilde{\sigma}^2 & 0 \\ 0 & \breve{\sigma}^2 \end{bmatrix}$ is the covariance matrix of the prediction residual signal of speech and noise, and where $\breve{\sigma}^2$ is the noise prediction residual power. $\mathbf{k}_n$ is denoted the Kalman gain which relies on the ratio of prediction error of speech and noise at frame n. The notation "$n|n-1$" means the prior estimate at acoustic frame n conditioned on the observation of all the previous frames $1, \ldots, n-1$. Therefore, $\mathbf{s}_{n|n-1}$ indicates the a priori estimate of the state vector while $\mathbf{s}_{n|n}$ indicates the a posteriori estimate. To determine the speech and noise model parameters, the time-frequency signal is segmented into overlapping modulation frames. For each frequency bin, a speech model $\left\{\tilde{\mathbf{b}},\, \tilde{\sigma}^2\right\}$ is estimated by applying autocorrelation LPC analysis to the modulation frame. However, the presence of noise will introduce bias in the LPC estimates, which will degrade the performance of the modulation domain Kalman filter. To alleviate

the effect of the noise, the MMSE enhancer described in Section 2.4 is applied to the noisy speech before the LPC model estimation in [66] . For the noise LPC model, a separate SNR-based VAD is applied to each frequency bin and a noise model, $\left\{\breve{\mathbf{b}},\, \breve{\sigma}^2\right\}$, is estimated during intervals where speech is absent. Unlike the SNR-based VADs reviewed in Section (2.2.1), where the SNR is calculated in each acoustic frame, the VAD is determined by the SNR in a modulation frame, which is computed as

$$\mathrm{SNR}_{\mathrm{mod}}(l,k) = 10\log_{10}\left(\frac{\sum_m |Z_l(m,k)|^2}{\sum_m |\widehat{W}_{l-1}(m,k)|^2}\right) \tag{2.25}$$

where $|\widehat{W}_{l-1}(m,k)|^2$ denotes the estimated noise modulation power spectrum of the previous modulation frame. If the $\mathrm{SNR}_{\mathrm{mod}}$ is larger than a preset threshold, the frequency bin is regarded speech present and vice verse. The noise power of the current modulation frame is estimated during speech absence using a forgetting factor κ , as given in (2.1)

$$|\widehat{W}_l(m,k)|^2 = \kappa|\widehat{W}_{l-1}(m,k)|^2 + (1-\kappa)\,|Z_l(m,k)|^2 \tag{2.26}$$

After the modulation power spectrum of noise is obtained, an ISTFT is applied for each modulation frame to get the corresponding autocorrelation coefficients, from which the LPC coefficients of noise can be estimated using Levinson-Durbin recursion [22].

The authors compared their enhancement algorithm with MMSE algorithm [7] and found that it consistently performed better in terms of Perceptual Evaluation of Speech Quality (PESQ) [69] and in terms of listener preference.

2.5.2. Modulation domain spectral subtraction

Apart from the modulation domain Kalman filter introduced in this subsection, the spectral subtraction method presented in Section 2.4 is applied in the modulation domain to estimate the modulation amplitude spectrum of the clean speech, $|\widehat{S}_l(m,k)|^2$, which is calculated by [65]

$$|\widehat{S}_l(m,k)|^2 = \left\{(|Z_l(m,k)|^r - \alpha|\widehat{W}_l(m,k)|^r)^{1/r}, \psi|\widehat{W}_l(m,k)|\right\}$$

where where $\alpha \geq 1$ and $0 \leq \psi \ll 1$ represent factors controlling the oversubtraction and the noise floor respectively as defined in Section 2.4. The noise modulation amplitude spectrum $|\widehat{W}_l(m,k)|$ is estimated using a method similar to that described in Section 2.5.1. The only difference is that $|\widehat{W}_l(m,k)|$, rather than $|\widehat{W}_l(m,k)|^2$, is updated using (2.26).

Using objective and subjective measures, it shows that the applying spectral subtraction in the modulation domain results in improved speech quality over the time-frequency domain spectral subtraction method [51] and the MMSE enhancer [7].

2.6. Enhancement Postprocessor

As explained in Chapter 1, although the time-frequency domain enhancement methods can improve the SNR of noisy speech signals, they also introduce spurious tonal artifacts including musical noise and speech distortion. One widely used method to remove the musical noise is by applying some form of post-processing to the output of the baseline enhancer or to the time-frequency gain function that it utilizes. The algorithm in [70] is proposed for post-processing the speech enhanced by spectral subtraction enhancer. It firstly classifies the spectrogram of the enhanced speech

into speech or musical-noise regions and for the musical-noise region, the spectral components corresponding to musical noise are identified and attenuated. In order to remove the musical noise in the subspace enhanced speech, a post-filtering method is proposed in [71] which applies masking thresholds estimated by first preprocessing the signal through spectral subtraction. This method is shown to be able to largely reduce the musical noise comparing with the spectral subtraction enhancer. Based on the analysis in the cepstral domain, the idea of [72] is to smooth the gain function in the cepstral domain, because the speech and unwanted noise artefacts is more decorrelated in this domain than in the STFT domain. Because the spectral peaks in the gain function caused by the artefacts are represented by higher cepstral coefficients, smoothing the higher coefficients can reduce their temporal dynamics. It is shown by subjective listening tests that this algorithm outperforms the enhancer in [35] which does not apply cepstral smoothing. Additionally, smoothing the enhancer gain function is used in [73] to attenuate musical noise in the frames in low SNR regions. The spectral gain function of an initial enhancer is smoothed by a low-pass filter and this algorithm is shown to give better performance after processing the gain function of the MMSE estimator in [74] and the MAP estimator in [55]. Under the assumption that the modulation domain LPC model of the clean speech is significantly different from that of the residual and musical noise, a post-processor using a modulation domain Kalman filter is proposed in [75], where the temporal dynamics of both speech and noise, are jointly modelled making use of a Kalman filter to give a optimal estimate of the clean speech amplitudes. The details of this post-processor will be given in Chapter 3.

2.7. Speech Quality Assessment

Speech quality is a judgment of a perceived multidimensional construct that is internal to the listener and is typically considered as a mapping between the desired and observed features of the speech signal. There are two types of speech quality assessment method. The first type is subjective methods, in which listeners give either an absolute ratings to one speech stimulus, or a preference to one speech stimulus over others. The most widely used quality scores obtained from a subjective experiment is referred as Mean Opinion Score (MOS) [76]. The MOS of the speech stimuli is rated by the listeners with five categories shown in Table 2.1. A numerical value (from 1 to 5) is assigned to each category. The score of the speech is obtained by averaging the values rated by all the listeners, which represents an overall perceptual quality of the degraded speech. Although the quality of a speech signal can be assessed in such a subjective experiment, it is time consuming and expensive when the number of speech stimuli is large. The second type of assessment is objective methods, which aim to overcome these issues by modeling the relationship between the desired and perceived characteristics of the signal algorithmically, without the use of listeners. Among the objective methods there are two main different types, and those which require a reference (clean) speech signal in addition to the received speech signal are referred to as intrusive methods, those which only use the received speech signal are referred to as non-intrusive methods [77]. In this thesis only intrusive objective measures will be used and the most popular of these are reviewed below.

The oldest and simplest type of intrusive measure are the SNR-based measures, which are calculated in the time domain and have low computational complexity. The classic SNR is calculated (in dB) as

MOS	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Table 2.1.: Categories of MOS [76].

$$\text{SNR} = 10\log_{10}\frac{\sum\limits_{n} s^2(n)}{\sum\limits_{n}\left\{s(n)-\hat{s}(n)\right\}^2} \tag{2.27}$$

where $\hat{s}(n)$ denotes the processed speech, which has been accurately time-aligned with the reference speech $s(n)$. The time alignment can be found by shifting the reference speech signal until the the correlation coefficient between $s(n)$ and $\hat{s}(n)$ is maximized. The calculation of the ratio of power in (2.27) is averaged over the entire signal. However, since the classic SNR is dominated by high energy portions of the speech signal, it does not reflect the overall speech quality. Therefore, there are variants of SNR based measures which have been proposed. In order to reflect the fluctuations of speech signal, a short-time version of SNR, which is referred to as segmental SNR (segSNR), is proposed. segSNR is calculated as the average of short-time SNR over each frame and is given by [78]

$$\text{SNR}_{\text{seg}} = \frac{1}{N}\sum_{m=0}^{N} 10\log_{10}\frac{\sum\limits_{t=Nm}^{Tm+T-1} s^2(t)}{\sum\limits_{t=Tm}^{Tm+T-1}\left\{s(t)-\hat{s}(t)\right\}^2} \tag{2.28}$$

where N is the total number of frames and typically the frame length is 10-20 ms. During the intervals of speech silence, segSNR can be negative because the speech energy is very small and these regions do not represent the contribution to the speech quality. Therefore, a VAD is often used before the calculation of the segSNR. In the

same vein, the frames with overly large or small energy do not reflect the quality well. As a result, a upper and lower bound are often set for segSNR, which is typically 35 and -10 dB. In addition, another widely used variation to the SNR measure, frequency-weighted SNR (fwSNR), is in the frequency domain and reflects the contribution of different frequency bands, which is computed as [79]

$$\text{fwSNR}_{\text{seg}} = \frac{10}{N}\sum_{n=0}^{N}\log_{10}\frac{\sum_{k=1}^{K}\omega_{n,k}\log_{10}\frac{|S_{n,k}|^2}{\left(|S_{n,k}|-|\hat{S}_{n,k}|\right)^2}}{\sum_{k=1}^{K}\omega_{n,k}} \tag{2.29}$$

where K represents the number of frequency bands and $\omega_{n,k}$ is the weight applied on the kth frequency band. $|\hat{S}_{n,k}|$ is the spectral amplitude of the degraded speech. The weights can be chosen in different ways and an example is to use the power of the reference speech amplitude with a power exponent smaller than 1 [79].

Most of the recent objective speech quality measures are perceptually motivated [69], among which the most popularly used in the evaluation of speech enhancement algorithms is an ITU standard (P. 862) PESQ [69]. The diagram of PESQ is given in Fig. 2.1. The aim of the PESQ measure is to model the signal processing in the peripheral auditory system and it is designed for using across a wide range of conditions. In PESQ, speech quality scores are calculated on a scale from -0.5 to 4.5 and a mapping function is then used to map the PESQ score to MOS. It has been reported that MOS mapped from PESQ has a correlation coefficient of 0.935 with the subjective MOS for a number of telecommunication relevant databases [80].

The Perceptual Objective Listening Quality Analysis (POLQA) metric is the successor of PESQ and is also an ITU standard (P. 863) measure [81], which is designed to overcome the weaknesses of PESQ such as the delays and sensitivity to time misalignment between the reference speech and processed speech. The major differences between POLQA and PESQ lie in the time alignment part and the perceptual

model. The time alignment process of POLQA is carried out before the comparison process. The output of this step is used for estimating the sampling frequency and delay compensation in the comparison process. For the perceptual model, POLQA uses both time and frequency masking which is significantly more accurate in imitating human perception of various distortions. The quality perception module in POLQA consists of a cognitive model which calculates the indicators of different acoustic characteristics such as frequency response, noise and room reverberation. The final POLQA score is determined by combining the different indicators which give a overall listening quality assessment. It is found that POLQA has been designed not only to provide an accurate MOS estimate for a large set of conditions specific to new codec and network technologies, but to also ensure higher accuracy for a wide range of degradations (e.g. various noise conditions).

In this thesis asses the enhancement quality will be assessed using segSNR and PESQ.the segSNR measure is used to assess the effect of the enhancement on the level of noise and PESQ to assess the speech quality. PESQ rather than POLQA has been used because the software for it was more readily available and because uncertainties in time-alignment are not an issue for the algorithms which are concerned with.

2.8. Conclusion

In this chapter, contributions in a number of fields related to single channel speech enhancement have been reviewed. The fields include noise estimation, subspace enhancement, time-frequency domain enhancement, modulation domain enhancement, postprocessor and speech quality assessment. In the following chapters of this thesis, a postprocessor based on the modulation domain Kalman filter present in Section

2.5.1 will be introduced in Chapter 3, a modulation subspace method based on the subspace method described in Section 2.3 will be introduced in Chapter 4, and two enhancers based on statistical models and the modulation domain Kalman filter will be introduced in Chapter 5.

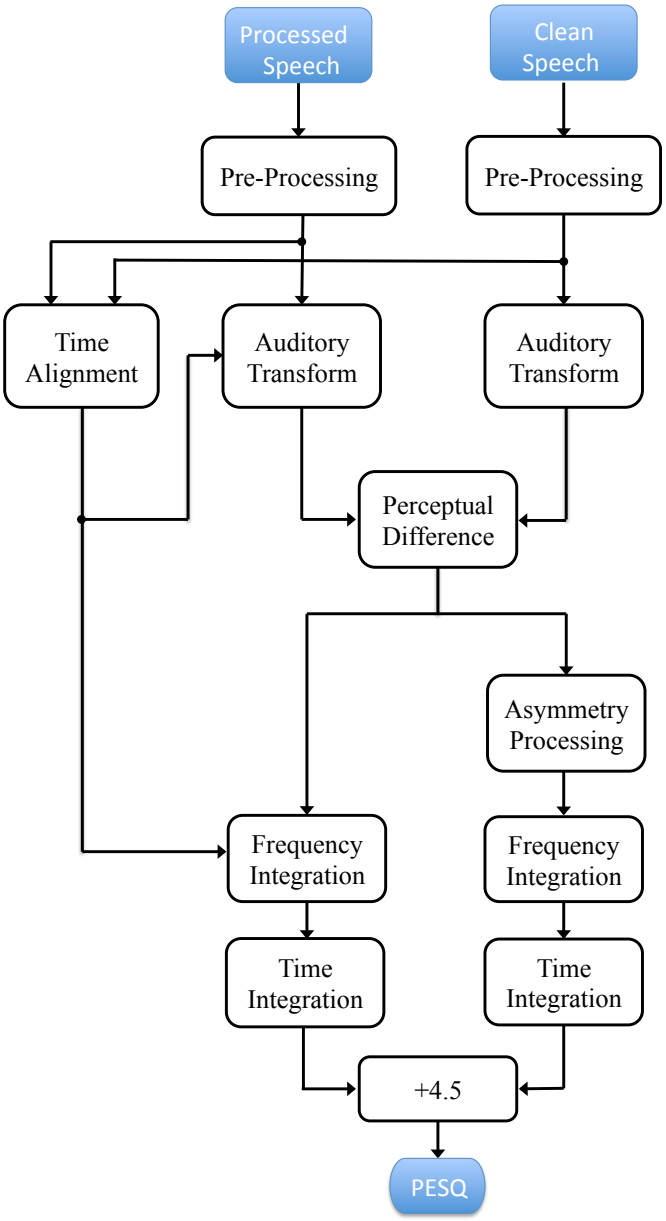

Figure 2.1.: Block diagram on the PESQ speech quality metric (diagram taken from [69]).

3. Modulation Domain Kalman Filtering

3.1. Introduction

As stated in Section 2.5, significant information in speech is carried by the modulation of spectral envelopes in addition to the envelopes themselves. There have been some speech enhancement algorithms extending models and techniques which were used in the time domain to the modulation domain. So and Paliwal have proposed applying the Kalman filter to the short-time modulation domain [66], the details of which has been given in Section 2.5. This Kalman filter incorporates autoregressive models for the temporal dynamics of the speech and noise spectral amplitudes in each frequency bin; these are estimated using Linear Predictive Coding (LPC) analysis. Because the clean speech and the noise in the MMSE enhanced speech have significantly different prediction characteristics in the modulation domain, in this chapter the use of a Kalman filter in the modulation domain will be introduced as a post-processor for speech that has been enhanced by an MMSE spectral amplitude algorithm [7]. Because the spectral amplitudes include a strong DC component, the gain of the corresponding LPC synthesis filter can be very high at low frequencies and therefore two alternative ways of constraining the low frequency gain in order

to improve the filter stability are proposed.

3.2. Kalman Filter Post-processing

The framework for our proposed speech enhancer is shown in Figure 3.1 and differs from that in [66] where the Kalman filter is applied not to the spectrum of the original noisy speech signal but rather to that of the output of an enhancer that implements the spectral amplitude MMSE algorithm from [7]. In our baseline system, denoted Modulation Domain Kalman filter post-processor (KFMD) in Section 3.2.4, the time-domain noisy speech, labelled $z(t)$ in Figure 3.1, is first transformed into the STFT domain and enhanced by the MMSE algorithm, from which the enhanced amplitude spectrum, $|Y(n,k)|$, can be obtained. Because the effect of the noise on the LPC estimation has been largely alleviated after the initial MMSE enhancement, the speech model is then estimated from the enhanced speech. The noise model is estimated from the MMSE enhanced spectral amplitudes using the method described in Section 2.5. The output from the Kalman filter is converted back to the amplitude domain, combined with the noisy phase spectrum, $\theta_{n,k}$, and passed through an ISTFT to create the output speech.

LPC is conventionally applied to a zero-mean time-domain signal [82] but in the modulation domain Kalman filter, it is applied to a positive-valued sequence of spectral amplitudes within each frequency bin. As will be shown in Section 3.2.1, when LPC analysis is applied to a signal that includes a strong DC component, the resultant synthesis filter can have a very high gain at low frequencies and the filter may, as a consequence, be close to instability. It has been found that this near-instability significantly degrades the quality of the output speech and thus in Section 3.2.2 and 3.2.3 two alternative ways of preventing it will be proposed.

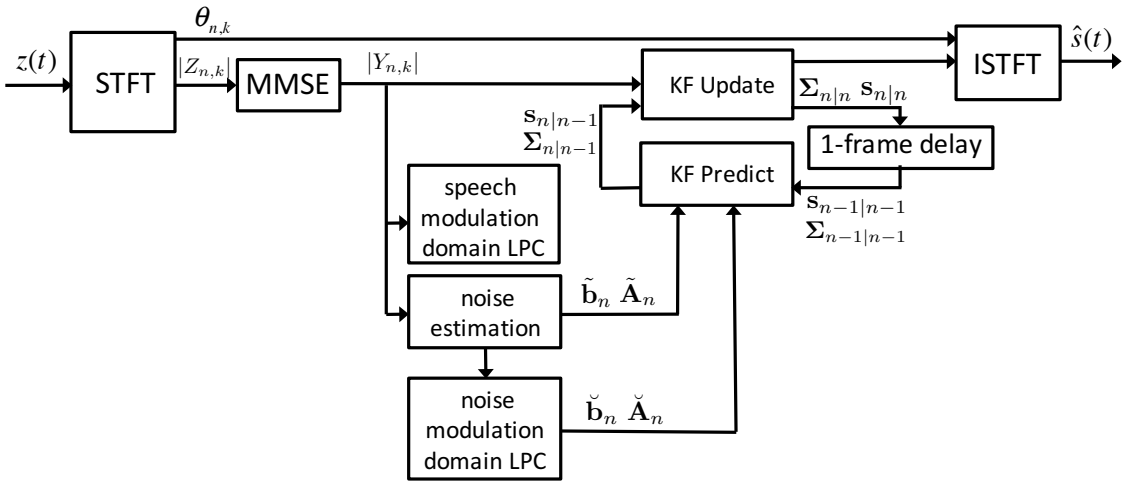

Figure 3.1.: Block diagram of KFMD algorithm

3.2.1. Effect of DC bias on LPC analysis

The speech amplitude spectrum $|S(n)|$ is generated in the modulation domain as the output of the modulation-domain LPC synthesis filter which is defined as

$$H(z) = \frac{1}{1 + \sum_{i=1}^{p} \tilde{b}_i z^{-i}} \tag{3.1}$$

where $\tilde{b}_i$ are the modulation-domain speech LPC coefficients defined in Section 2.5. Here the effect of a strong DC component on the results of LPC analysis is analyzed. Suppose first that the temporal envelope of the speech power spectrum $|S(n)|$ has zero mean and that the speech LPC coefficient vector, $\tilde{\mathbf{b}}$, for a frame of length L is determined from the Yule-Walker equations

$$\tilde{\mathbf{b}} = -\mathbf{R}^{-1}\mathbf{g} \tag{3.2}$$

where the elements of the autocorrelation matrix, $\mathbf{R}$, are given by $R_{i,j} = \frac{1}{L}\sum_n |S(n-i)||S(n-j)|$ for $1 \leq i, j \leq p$ and the elements of $\mathbf{g}$ are $g_i = R_{i,0}$. The DC gain of

the synthesis filter $H(z)$, obtained by setting $z = 1$ in (3.1), is given by

$$G_H = \frac{1}{1 + \mathbf{o}^T \tilde{\mathbf{b}}} \tag{3.3}$$

where $\mathbf{o} = [1\ 1\ \cdots\ 1]^T$ is a p-dimensional vector of ones. For a filter, a very small DC gain indicates that the filter have zeros which are very close to the unit circle. On the other hand, a very large DC gain shows that the filter have poles which are very close to the unit circle, therefore in this case, the filter has very large gains at low frequencies and is near instability.

If now a DC component, d_s, is added to each $|S(n)|$, the effect is to add d_s^2 onto each $R_{i,j}$ and the new LPC coefficients, $\tilde{\mathbf{b}}'$, are given by

$$\begin{aligned} \tilde{\mathbf{b}}' &= -\left(\mathbf{R} + d_s^2 \mathbf{o}\mathbf{o}^T\right)^{-1} \left(\mathbf{g} + d_s^2 \mathbf{o}\right) \\ &= -\left(\mathbf{R}^{-1} - \frac{d_s^2 \mathbf{R}^{-1}\mathbf{o}\mathbf{o}^T\mathbf{R}^{-1}}{1 + d_s^2 \mathbf{o}^T\mathbf{R}^{-1}\mathbf{o}}\right) \left(\mathbf{g} + d_s^2 \mathbf{o}\right) \end{aligned}$$

where the second line follows from the Matrix Inversion Lemma [83]. Writing

$$x = d_s^2 \mathbf{o}^T \mathbf{R}^{-1} \mathbf{o} \tag{3.4}$$

then it can be obtained that

$$\mathbf{o}^T \tilde{\mathbf{b}}' = \frac{-\mathbf{o}^T\mathbf{R}^{-1}\mathbf{g} - x}{1 + x} = \frac{\mathbf{o}^T\tilde{\mathbf{b}} - x}{1 + x}.$$

Thus the DC gain of the new synthesis filter is

$$\frac{1}{1 + \mathbf{o}^T \tilde{\mathbf{b}}'} = \frac{1 + x}{1 + \mathbf{o}^T \tilde{\mathbf{b}}} \tag{3.5}$$

From (3.5) it can be seen that the DC gain of the synthesis filter has been multiplied by $1 + x$ where x, defined by (3.4), is proportional to the power ratio of the DC

and AC components of $|S(n)|$. If this ratio is large, the low frequency gain of the LPC synthesis filter can become very high which results in near instability and poor prediction. Accordingly, in the following sections two alternative methods of limiting the low frequency gain of the LPC synthesis filter are proposed.

3.2.2. Method 1: Bandwidth Expansion

The technique of bandwidth expansion is widely used in coding algorithms to reduce the peak gain and improve the stability of an LPC synthesis filter [84]. If a modified set of LPC coefficient is defined by $\dot{b}_i = c^i b_i$, for some constant $c < 1$, then the poles of the synthesis filter are all multiplied by c. This can be proved by substituting b_i with $\dot{b}_i$ in (3.1) and it is equivalent to replacing z with $\frac{z}{c}$. This moves the poles away from the unit circle thereby reducing the gain of the corresponding frequency domain peaks and improving the stability of the filter. In Section 3.2.4 the effect of using this revised set of LPC coefficients, $\tilde{\mathbf{b}}_1$, in the Kalman filter of Figure 3.1 (denoted the "BKFMD" algorithm) will be evaluated and find that it results in a consistent improvement in performance.

3.2.3. Method 2: Constrained DC gain

Although the bandwidth expansion approach is effective in limiting the low frequency gain of the synthesis filter, it also modifies the filter response at higher frequencies thereby destroying its optimality. This effect can be seen in 3.2, where the LPC analysis is applied on a modulation frame with strong speech power. An alternative approach is to constrain the DC gain of the synthesis filter to a predetermined value and determine the optimum LPC coefficients subject to this constraint. As noted in Section 3.2.1, the DC gain of the LPC synthesis filter is given by G_H in

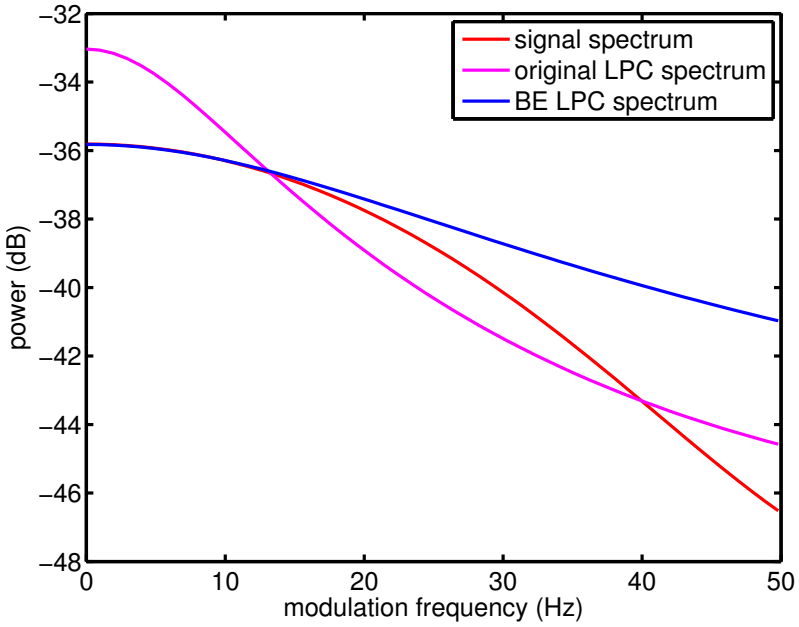

Figure 3.2.: Smoothed power spectrums of the modulation domain signal, original LPC filter, the bandwidth expansion (BE) LPC filter. The LPC spectrums and signal spectrum are calculated from the same modulation frame and $c = 0.7$.

(3.3) and $G_H = G_0$ can be forced by imposing the constraint

$$\mathbf{o}^T\tilde{\mathbf{b}} = \frac{1 - G_0}{G_0} \triangleq \beta_G > -1.$$

The average prediction error energy in the analysis frame is given by

$$E = \frac{1}{L}\sum_n \left\{ |S(n)| + \sum_{i=1}^{p} b_i |S(n-i)| \right\}^2$$

and E is going to be minimized subject to the constraint $\mathbf{o}^T\tilde{\mathbf{b}} = \beta_G$. Using a Lagrange multiplier, λ, the solution, $\tilde{\mathbf{b}}_2$ to this constrained optimization problem is

obtained by solving the $p+1$ equations

$$\begin{aligned} \frac{d}{da_i}\left(E + \lambda \mathbf{o}^T \tilde{\mathbf{b}}_2\right) &= 0 \\ \mathbf{o}^T \tilde{\mathbf{b}}_2 &= \beta_G \end{aligned}$$

and the solution is

$$\begin{pmatrix} 0.5\lambda \\ \tilde{\mathbf{b}}_2 \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{o}^T \\ \mathbf{o} & \mathbf{R} \end{pmatrix}^{-1} \begin{pmatrix} \beta_G \\ -\mathbf{g} \end{pmatrix} \tag{3.6}$$

where $\mathbf{R}$, $\mathbf{g}$ and $\mathbf{o}$ are as defined in Section 3.6. As shown in Figure 3.3, this revised LPC model can lower the filter gains at low modulation frequencies when keeping the gains at high modulation frequencies closed to the unconstrained LPC model. In Section 3.2.4 the effect of using this set of LPC coefficients, $\tilde{\mathbf{b}}_2$, in the Kalman filter of Figure 3.1 (denoted the "CKFMD" algorithm) will be evaluated and find that it results in a consistent improvement in performance both over the KFMD algorithm, which uses the unconstrained filter coefficients, $\tilde{\mathbf{b}}$, and also over the BKFMD algorithm which uses the bandwidth expanded coefficients, $\tilde{\mathbf{b}}_1$.

3.2.4. Evaluation

In this section, the performance of the baseline MMSE enhancer [85] is compared with that of the three algorithms that incorporate a Kalman filter post-processor. The KFMD algorithm which uses an unconstrained speech model, the BKFMD algorithm incorporates the bandwidth expansion from 3.2.2 while the CKFMD algorithm uses the constrained filter from Section 3.2.3. In our experiments, the core test set from the TIMIT database is used and the speech is corrupted by 'white' and 'factory1' noise from the RSG-10 database [25] at −5, 0, 5, 10, 15, and 20 dB SNR. The algorithm parameters were determined by optimizing performance with

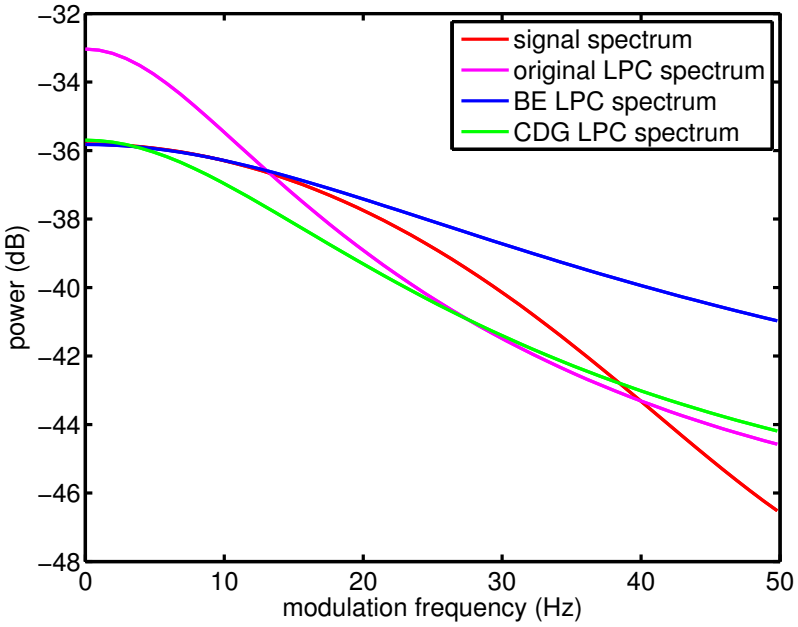

Figure 3.3.: Smoothed power spectrums of the modulation domain signal, original LPC filter, the LPC filter with a constrained DC gain (CDG). The LPC spectrums and signal spectrum are calculated from the same modulation frame and $\beta_G = -0.8$ in (3.6).

respect to PESQ on the development set described in Section 1.4.1.1. The parameter settings have been listed in Table 3.1.

Using the new LPC models, the performance of the speech enhancers is evaluated using both segSNR and PESQ measures. In all cases, the measures are averaged over all the sentences in the TIMIT core test set. Figures 3.4 and 3.5 show how the average segSNR varies with global SNR for white noise and factory noise for the unenhanced speech, the baseline MMSE enhancer and the three Kalman filter postprocessing algorithms presented in this subsection. It can be seen that at high SNRs, all the algorithms have very similar performance. However at 0 dB SNR the KFMD provides an approximate 2 dB improvement in segSNR over MMSE enhancement and the BKFMD and CKFMD algorithms give an additional 0.5 and 1.5 dB improvement respectively. The PESQ results shown in Figures 3.6 and 3.7 broadly

Parameter	Settings
Sampling frequency	8 kHz
Acoustic frame length	16 ms
Acoustic frame increment	4 ms
Modulation frame length	64 ms
Modulation frame increment	16 ms
Analysis-synthesis window	Hamming window
Speech LPC model order p	3
Noise LPC model order q	4
Bandwidth expansion coefficient c	0.7
Constrained DC gain β_G	-0.8

Table 3.1.: Parameters settings in experiments.

mirror the segSNR results although the post-processing gives an improvement in PESQ even at high SNRs. For both noise types, the constrained Kalman filter postprocessor (CKFMD) gives a PESQ improvement of > 0.2 over a wide range of SNRs. The consistent improvements in performance for both the stationary noise (white noise) and non-stationary noise (factory noise) show that incorporating the dynamical modelling of noise is beneficial for noise reduction for both types of noises.

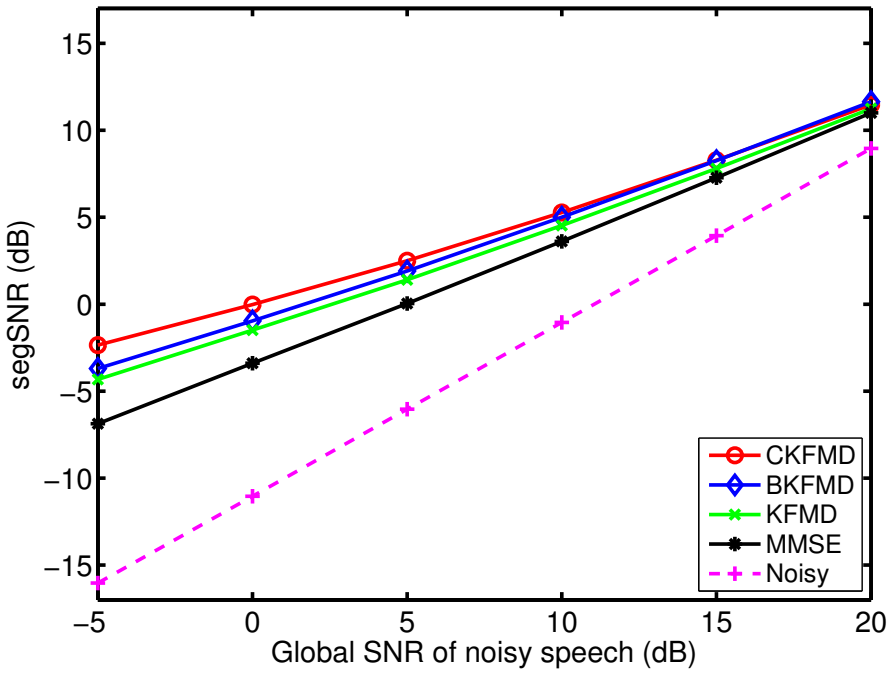

Figure 3.4.: Average segSNR values comparing different algorithms, where speech signals are corrupted by white noise at different SNR levels.

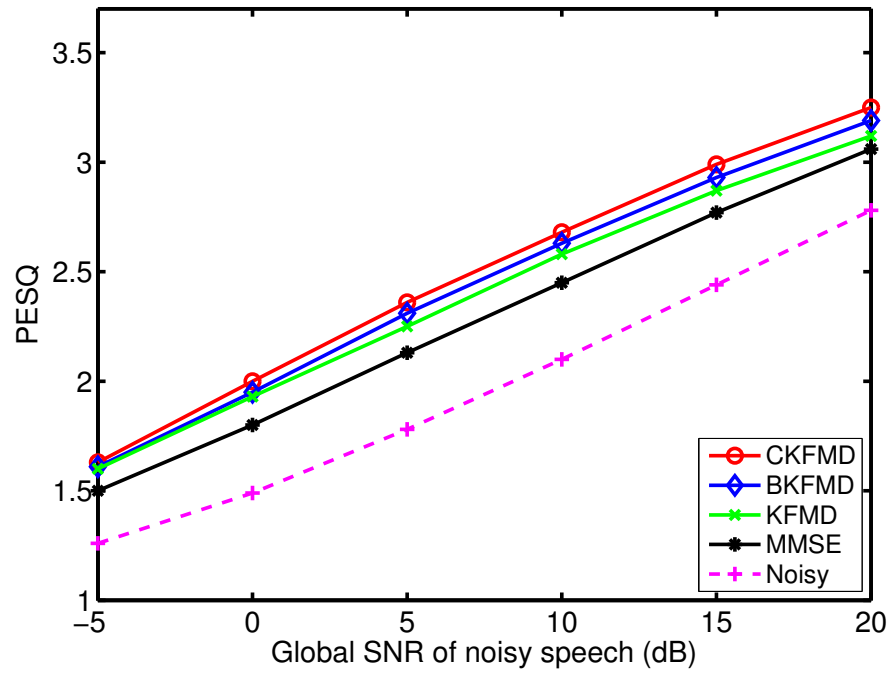

Figure 3.6.: Average PESQ values comparing different algorithms, where speech signals are corrupted by white noise at different SNR levels.

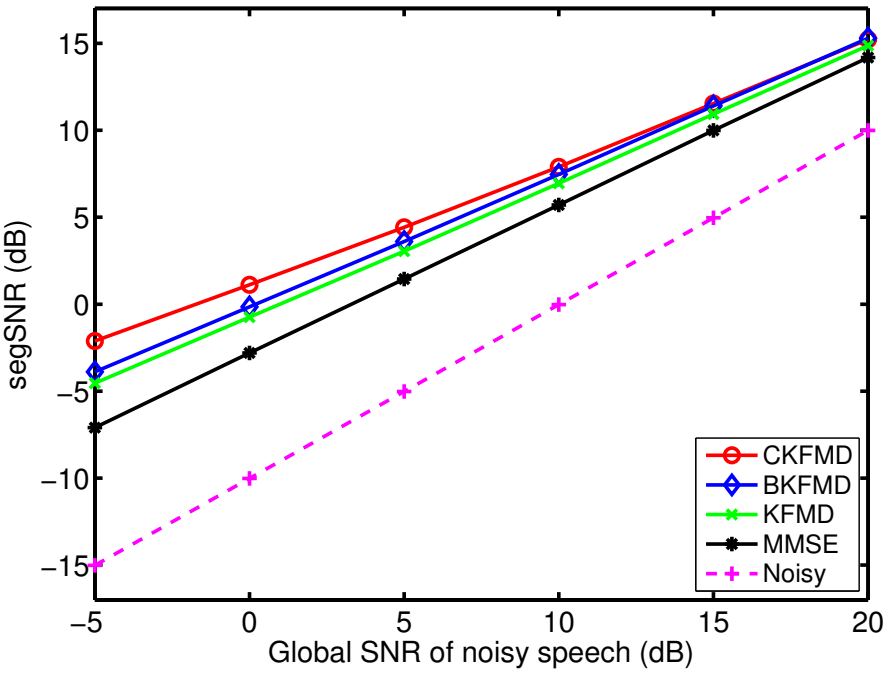

Figure 3.5.: Average segSNR values comparing different algorithms, where speech signals are corrupted by factory noise at different SNR levels.

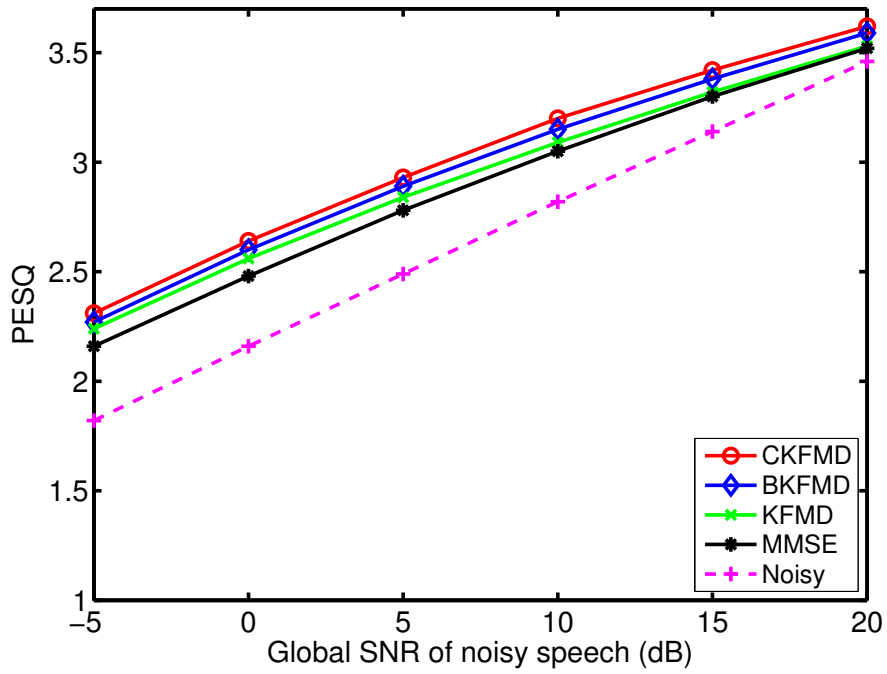

Figure 3.7.: Average PESQ values comparing different algorithms, where speech signals are corrupted by factory noise at different SNR levels.

3.3. GMM Kalman filter

In the conventional Kalman filter introduced above, the prediction residual signal of both speech and noise are assumed Gaussian distributed. However, after processing noisy speech by an MMSE enhancer, most of the stationary noise has been removed leaving behind some residual noise together with musical noise artefacts, especially where the input noise power was high [12], as shown in Figure 1.3 in Chapter 1. As described in Section 1.2.2, because the musical noise is characterized by isolated spectral peaks in the spectrogram, it is difficult to predict in the modulation domain. As a result, the prediction errors associated with the musical noise may be very large, and the overall distribution of the prediction errors of the noise in the enhanced speech does not follow a Gaussian distribution. To illustrate this, in Figure 3.8 the distribution of the normalized prediction error of the spectral amplitude errors in the MMSE enhanced speech in all frequency bins together with a fitted single Gaussian distribution (in red) and a 3-mixture Gaussian Mixture Model (GMM) (in green) is shown. The histogram shows the distribution over all time-frequency bins using the TIMIT core test set corrupted by additive car noise at SNRs between –10 and +15 dB using the framing parameters from Section 3.3.3. The estimated noise amplitude trajectory in each frequency bin is represented by an autoregressive model and the model parameters (LPC coefficients) are estimated in the corresponding modulation frame. To obtain a general distribution that is independent of the noise level, the normalized residual rather than the residual itself is modeled so that the GMM parameters are independent of the speech and noise amplitudes. The residual signals are normalized by the RMS power of the noise predictor residual in the corresponding modulation frame. The figure shows that the overall prediction residual signal is not zero mean and does not follow a Gaussian distribution.

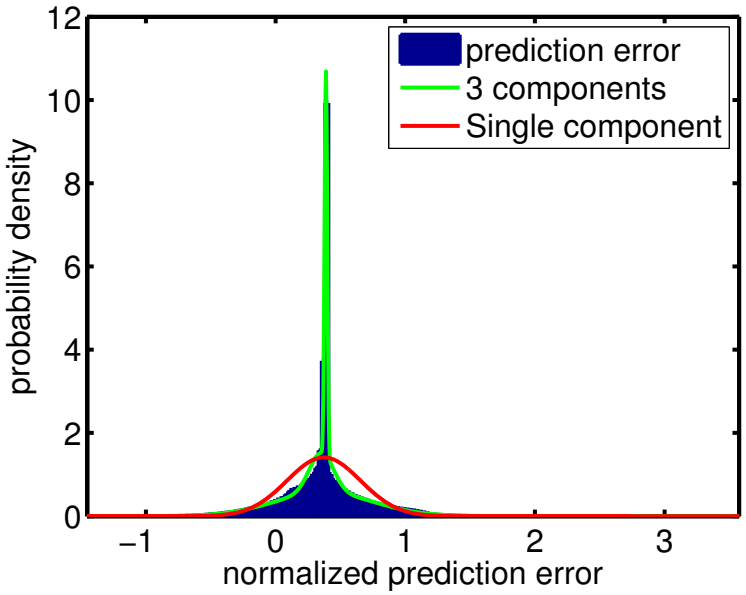

Figure 3.8.: Distribution of the normalized prediction error of the noise spectral amplitudes in MMSE-enhanced speech. The prediction errors are normalized by the RMS power of the noise predictor residual in the corresponding modulation frame.

3.3.1. Derivation of GMM Kalman filter

Based on the empirical prediction errors, the conventional colored noise KF has been extended to incorporate a GMM noise distribution. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is used to denote a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and use $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ for its probability density at $\mathbf{x}$. The advantage of Gaussian mixture model is twofold: first, it is flexible to fit various distributions; second, the posterior distribution of the estimation is still Gaussian mixtures whose parameters are efficient to compute.

The diagram of the proposed algorithm is shown in Figure 3.9. Following time-frequency domain enhancement in the block marked MMSE, the spectral amplitude of the STFT at time frame n and frequency bin k is given by $|Y_{n,k}| = |S_{n,k}| + |W_{n,k}|$, where the amplitudes $|W_{n,k}|$ here represents the "noise" arising from a combination

of acoustic noise and the enhancement artefacts. The output from the Kalman filter $|\hat{S}_{n,k}|$ is combined with the noisy phase spectrum $\theta_{n,k}$ and passed through an ISTFT to create the output speech $\hat{s}(t)$. In this and the next subsection the derivation of the GMM Kalman filter and the parameter update procedure will be given. Because each frequency bin, k, is processed independently and for clarity, the frequency index will be omitted below.

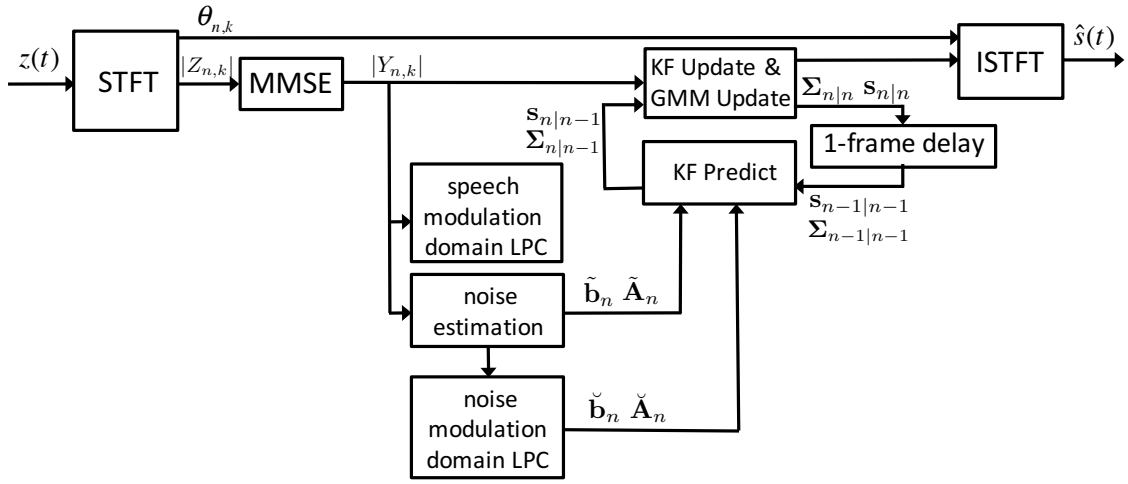

Figure 3.9.: Diagram of the proposed GMM KF algorithm

The system model and the Kalman filter equations are given in Section 2.5.1, the prediction residuals are represented as a 2-element vector $\mathbf{e}_n$ with a Gaussian mixture distribution of J mixtures as

$$\mathbf{e}_n \sim \sum_{j=1}^{J} \epsilon_n^{(j)} \mathcal{N}(\boldsymbol{\mu}_n^{(j)}, \boldsymbol{\Sigma}_n^{(j)}) \tag{3.7}$$

where $\epsilon_n^{(j)}$ is the weight of each mixture j and the sum of $\epsilon_n^{(j)}$ over all J mixtures satisfies $\sum_{j=1}^{J} \epsilon_n^{(j)} = 1$. $\boldsymbol{\mu}_n^{(j)}$ and $\boldsymbol{\Sigma}_n^{(j)}$ are the mean vector and covariance matrix of each mixture. As in a conventional Kalman filter, the augmented state vector $\mathbf{s}_n$ at time $n-1$ based on observations up to time $n-1$ is assumed to be Gaussian distributed $\mathbf{s}_{n-1} \sim \mathcal{N}(\mathbf{s}_{n-1|n-1}, \boldsymbol{\Sigma}_{n-1|n-1})$. Following the time update, the distribution of $\mathbf{s}_{n|n-1}$

becomes a Gaussian mixture $\sum_j \epsilon_{n-1}^{(j)} \mathcal{N}(\mathbf{s}_{n|n-1}^{(j)}, \mathbf{\Sigma}_{n|n-1}^{(j)})$ where

$$\mathbf{s}_{n|n-1}^{(j)} = \mathbf{A}_{n-1}\mathbf{s}_{n-1|n-1} + \mathbf{D}_1\boldsymbol{\mu}_{n-1}^{(j)}$$
$$\mathbf{\Sigma}_{n|n-1}^{(j)} = \mathbf{A}_{n-1}\mathbf{\Sigma}_{n-1|n-1}\mathbf{A}_{n-1}^T + \mathbf{D}_1\mathbf{Q}_{n-1}^{(j)}\mathbf{D}_1^T.$$

Applying the observation constraint, $\mathbf{D}_2^T\mathbf{s}_n = |Y_n|$, changes the Gaussian mixture parameters as follows [83]

$$\mathbf{k}_n^{(j)} = \mathbf{\Sigma}_{n|n-1}^{(j)}\mathbf{D}_2(\mathbf{D}_2^T\mathbf{\Sigma}_{n|n-1}^{(j)}\mathbf{D}_2)^{-1} \tag{3.8}$$
$$\mathbf{s}_{n|n}^{(j)} = \mathbf{s}_{n|n-1}^{(j)} + \mathbf{k}_n^{(j)}(|Y_n| - \mathbf{s}_{n|n-1}^{(j)}) \tag{3.9}$$
$$\mathbf{\Sigma}_{n|n}^{(j)} = \mathbf{\Sigma}_{n|n-1}^{(j)} - \mathbf{k}_n^{(j)}\mathbf{D}_2^T\mathbf{\Sigma}_{n|n-1}^{(j)}. \tag{3.10}$$

Finally, the GMM is collapsed into a single Gaussian for the estimation of the state vector at time n, by calculating the overall mean and covariance matrix of the posterior Gaussian mixture [86].

$$\pi_n^{(j)} = \frac{\epsilon_{n-1}^{(j)}\mathcal{N}(|Y_n|; \mathbf{D_2}^T\mathbf{s}_{n|n-1}^{(j)}, \mathbf{D}_2^T\mathbf{\Sigma}_{n|n-1}^{(j)}\mathbf{D}_2)}{\sum_j \epsilon_{n-1}^{(j)}\mathcal{N}(|Y_n|; \mathbf{D}_2^T\mathbf{s}_{n|n-1}^{(j)}, \mathbf{D}_2^T\mathbf{\Sigma}_{n|n-1}^{(j)}\mathbf{D}_2)} \tag{3.11}$$
$$\mathbf{s}_{n|n} = \sum_{j=1}^{J} \pi_n^{(j)}\mathbf{s}_{n|n}^{(j)} \tag{3.12}$$
$$\mathbf{\Sigma}_{n|n} = \sum_{j=1}^{J} \pi_n^{(j)}(\mathbf{\Sigma}_{n|n}^{(j)} + \mathbf{s}_{n|n}^{(j)}(\mathbf{s}_{n|n}^{(j)})^T) - \mathbf{s}_{n|n}\mathbf{s}_{n|n}^T. \tag{3.13}$$

The quantity $\pi_n^{(j)}$ in (3.11) represents the posterior probability that $\mathbf{s}_n$ belongs to mixture j.

Thus the new Kalman filter can be used to process the residual noise in the MMSE enhanced speech because the GMM can be used to model the spectral amplitude errors in the enhanced speech. In this work, the initial GMM parameters are trained

on speech sentences from the training set of TIMIT database using expectation maximization algorithm [86]. A method for updating the parameters will be present in the following subsection.

3.3.2. Update of parameters

The spectral amplitudes, $|Y_{n,k}|$ are divided into overlapping modulation frames and autocorrelation LPC analysis [22] is performed in each modulation frame to obtain a vector of modulation-domain LPC coefficients, $\tilde{\mathbf{b}}$, and a residual power $\tilde{\sigma}^2$. To obtain the corresponding noise coefficients, the sequence of spectral amplitudes, $|Y_{n,k}|$, is passed through a noise power spectrum estimator [43] before performing LPC analysis to obtain the noise predictor coefficients, $\breve{\mathbf{b}}$, and the residual power $\breve{\sigma}^2$.

Within the noise GMM, (3.7), the speech residual component $\tilde{e}_n \sim \mathcal{N}(0, \tilde{\sigma_n}^2)$ is identical in all mixture components but the normalized noise residual $v_n = \breve{e}_n/\breve{\sigma}_n$ is modeled as a Gaussian mixture

$$v_n \sim \sum_j \epsilon_n^{(j)} \mathcal{N}(m_n^{(j)}, \rho_n^{2(j)}). \tag{3.14}$$

As mentioned above, the normalized residual rather than the residual itself is modeled so that the GMM parameters are independent of the speech and noise amplitudes.

In order to update the GMM parameters the noise predictor coefficients, $\breve{\mathbf{b}}$, from the current modulation frame are applied to the sequence of estimated noise spectral amplitudes to obtain a noise prediction error $v_n\breve{\sigma}_n$ for each acoustic frame n. The probability that v_n comes from mixture j is given by

$$p_n^{(j)} = \frac{\epsilon_{n-1}^{(j)} \mathcal{N}(v_n; m_{n-1}^{(j)}, \rho_{n-1}^{2(j)})}{\sum_j \epsilon_{n-1}^{(j)} \mathcal{N}(v_n; m_{n-1}^{(j)}, \rho_{n-1}^{2(j)})}. \tag{3.15}$$

Because now the probability of the mixture given the observation error is known, the statistics accumulated from the previous frames can be update in the current frame. The statistics include the effective number of observations ($O^{(j)}$), the sum of the observations ($\Psi^{(j)}$) and the sum of the squared observations ($T^{(j)}$) as $O_n^{(j)} = p_n^{(j)} + \kappa O_{n-1}^{(j)}$, $\Psi_n^{(j)} = p_n^{(j)} v_n + \kappa \Psi_{n-1}^{(j)}$ and $T_n^{(j)} = p_n^{(j)} v_n^2 + \kappa T_{n-1}^{(j)}$, where κ is a forgetting factor. The parameters, $m_n^{(j)}$, $\rho_n^{2(j)}$ and $\epsilon_n^{(j)}$, in (3.14) can now be updated adaptively as [86]

$$m_n^{(j)} = \Psi_n^{(j)} / O_n^{(j)} \tag{3.16}$$

$$\rho_n^{2(j)} = T_n^{(j)} / O_n^{(j)} - m_n^{2(j)} \tag{3.17}$$

$$\epsilon_n^{(j)} = \frac{O_n^{(j)}}{\sum_j O_n^{(j)}} = (1-\kappa)\, O_n^{(j)} \tag{3.18}$$

To initialize the model, a GMM with parameters $m_0^{(j)}$, $\rho_0^{2(j)}$ and $\epsilon_0^{(j)}$ is trained offline on a large amount of data and set $O_0^{(j)} = m_0^{(j)}/(1-\kappa)$, $\Psi_0^{(j)} = m_0^{(j)} O_0^{(j)}$ and $T_0^{(j)} = (\rho_0^{2(j)} + m_0^{2(j)}) O_0^{(j)}$. To ensure stability of the update procedure, lower bounds on $p^{(j)}$ and $\rho^{2(j)}$ are imposed to prevent them from becoming zero.

3.3.3. Evaluation

In this subsection, the performance of the proposed Kalman filter post-processor with a GMM noise model (KFGM) is compared with the baseline MMSE enhancer from [7] and the KFMD from Section 3.2. The constrained LPC model introduced in 3.2.3 is also combined with the algorithm and the resulting algorithm is referred to as CKFGM. The initial GMM parameters are trained using a subset in the training set of the TIMIT database comprising 500 sentences and using speech corrupted

by white noise. The remaining algorithm parameters were chosen to optimize the performance of the algorithms, with respect to PESQ, on the development set and their values are listed in Table 3.2. In the experiments, the core test set from the TIMIT database (details in Chapter 2) is used and the speech is corrupted by the 'factory2' noise from the RSG-10 database [25] and 'street' noise from the ITU-T test signals database [87] at $-10, -1, 0, 5, 10$ and 15 dB global SNR. The reason street noise, rather than white noise that is used in Section 3.2.4, is used is that in this section non-stationary colored noises are more appropriate to evaluate the performance of the GMM noise model that is incorporated in the modulation domain Kalman filter postprocessor.

Parameter	Settings
Sampling frequency	8 kHz
Acoustic frame length	16 ms
Acoustic frame increment	4 ms
Modulation frame length	128 ms
Modulation frame increment	16 ms
Analysis-synthesis window	Hamming window
Number of mixtures J	3
Speech LPC model order p	3
Noise LPC model order q	4
Forgetting factor κ	0.9

Table 3.2.: Parameter settings in experiments.

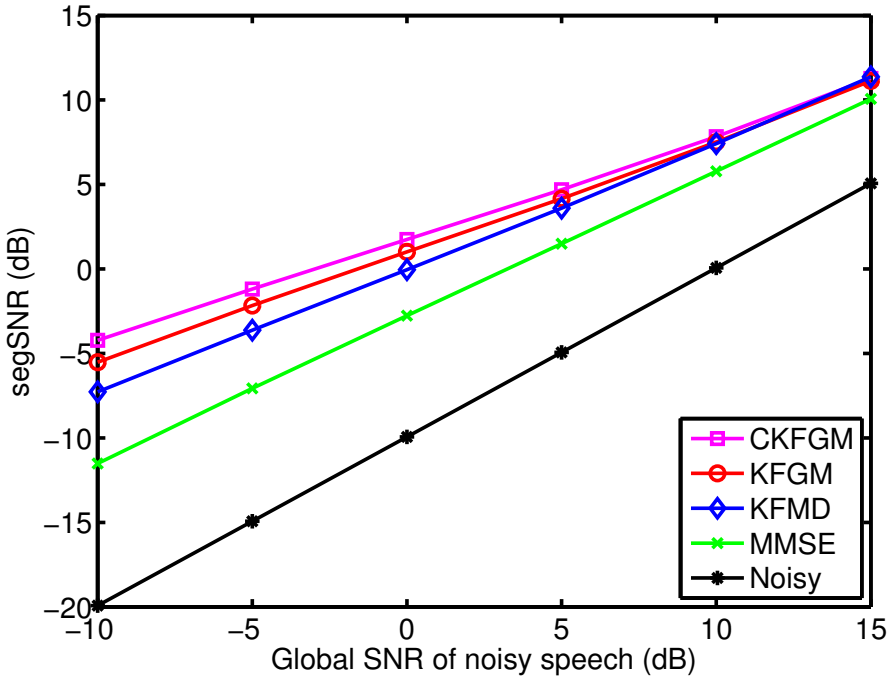

Figure 3.10.: Average segmental SNR of enhanced speech after processing by four algorithms versus the global SNR of the input speech corrupted by factory noise (CKFGM: proposed Kalman filter post-processor with a constrained LPC model and a Gaussian Mixture noise model; KFGM: proposed KFGM algorithm; KFMD: KFMD algorithm from [75]; MMSE: MMSE enhancer from [7]).

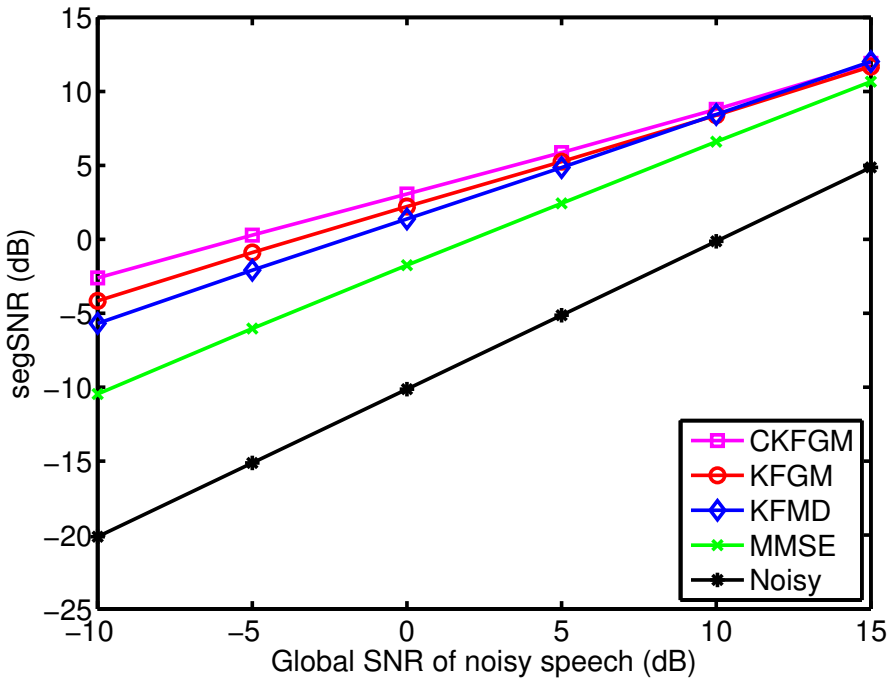

Figure 3.11.: Average segmental SNR of enhanced speech after processing by four algorithms versus the global SNR of the input speech corrupted by street noise.

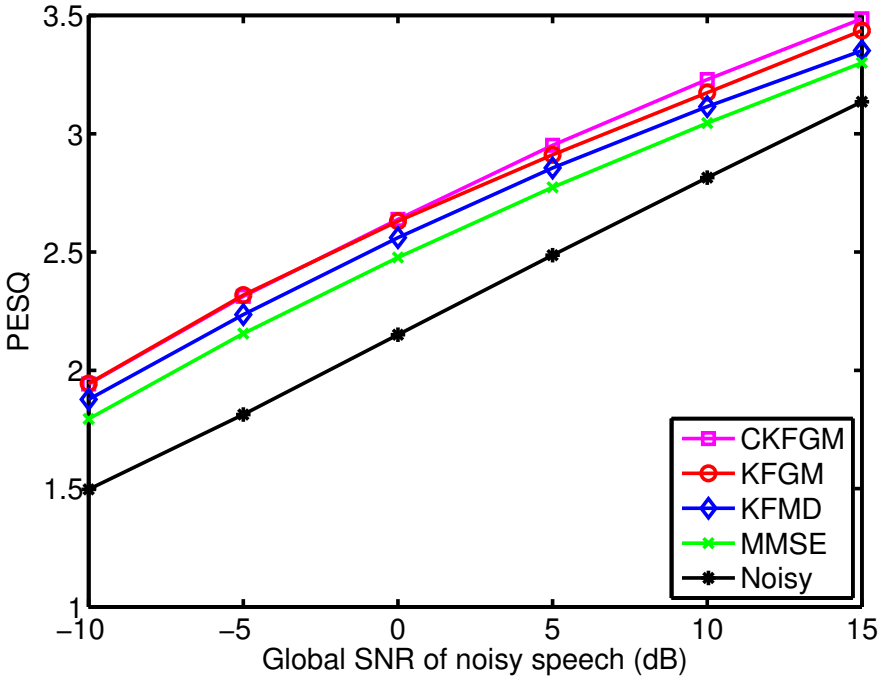

Figure 3.12.: Average PESQ quality of enhanced speech after processing by four algorithms versus the global SNR of the input speech corrupted by factory noise.

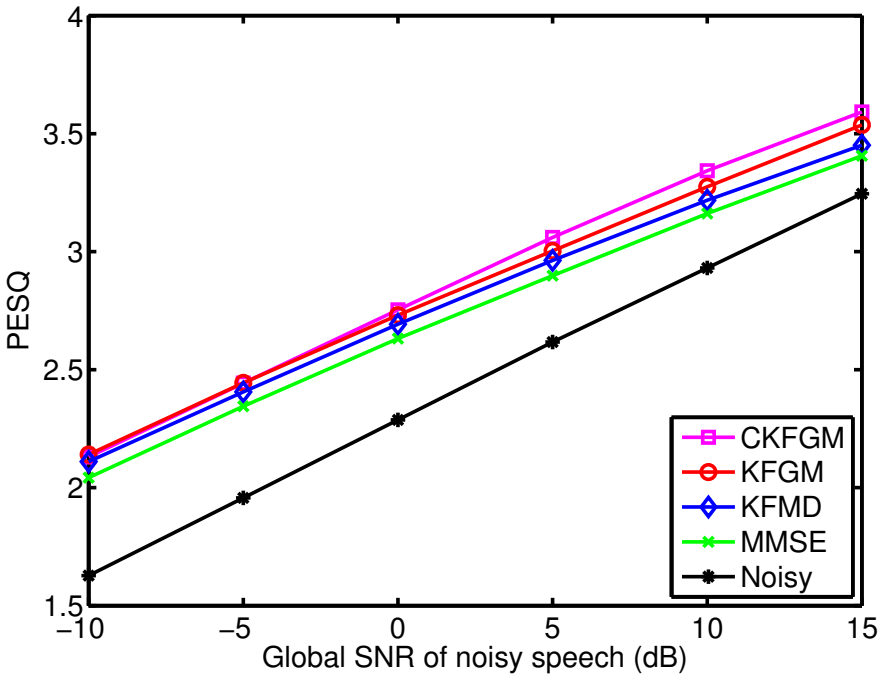

Figure 3.13.: Average PESQ quality of enhanced speech after processing by four algorithms versus the global SNR of the input speech corrupted by street noise.

The performance of the algorithms is evaluated using both segmental SNR (segSNR) and the Perceptual Evaluation of Speech Quality (PESQ) measure. All the measurement values are averaged over the 192 sentences in the TIMIT core test set. The average segSNR for the corrupted speech, baseline MMSE enhancer, the KFMD algorithm, the proposed KFGM algorithm and the KFGM algorithm using the constrained LPC model derived in Section 3.2.3 (CKFGM) is shown for factory noise in Figure 3.10 as a function of the global SNR of the noisy speech. It can be seen that at 15 dB global SNR all the algorithms give the same improvement in segSNR of about 5 dB. However, at 0 dB global SNR the KFGM algorithm outperforms both reference algorithms by about 1 dB and 3 dB respectively, and CKFGM algorithm gives an additional 1 dB improvement. The equivalent graphs for street noise are shown in Figure 3.11. It can be seen the overall trend in the results is the same. The corresponding graphs for PESQ are shown in Figure 3.12 for factory noise and in Figure 3.13 for street noise. In Figures 3.12 and 3.13, the average PESQ scores mirror the results seen for the segSNR. However, at high SNRs the KFGM algorithm is also able to improve the PESQ, and an improvement of approximately 0.1 and 0.15 over the algorithm and MMSE enhancer respectively can be obtained over a wide range of SNRs. By using the constrained LPC model it can get even better performance at high SNRs as it can be seen that the CKFGM algorithm outperform the KFGM algorithm by about 0.1 PESQ at 15 dB SNR. This shows that incorporating a better speech LPC model can also lead to better performance for KFGM algorithm. In addition, informal listening tests also suggest that the proposed post-processing methods is able to reduce the musical noise introduced by the MMSE enhancer.

3.4. Conclusion

In this chapter two different methods of post-processing the output of an MMSE spectral amplitude speech enhancer by using a Kalman filter in the modulation domain have been proposed. Firstly, different speech LPC models in each modulation frame is introduced and it is shown that the post-processors based on the LPC models give consistent improvements over the MMSE enhancer in both segSNR and PESQ, among which the best method, which performs LPC analysis with a constrained DC gain, improves PESQ scores by at least 0.2 over a wide range of SNRs. Secondly, a post-processor in the modulation domain using a GMM for modeling prediction error of the noise in the output spectral amplitude of MMSE enhancer is introduced. The derivation of a Kalman filter that incorporates a GMM noise model has been given and a method for adaptively updating the GMM parameters has also been presented. The proposed post-processor has been evaluated using segSNR and PESQ and shown that the proposed method results in consistently improved performance when compared to both the baseline MMSE enhancer and a modulation-domain Kalman filter post-processor. The improvement in segSNR is over 3 dB at a global SNR of 0 dB while the PESQ score is increased by about 0.15 across a wide range of input global SNRs. The results show that a GMM is preferable to a single Gaussian model in modelling the prediction residual of the spectral amplitudes of the musical noise under non-stationary colored noise conditions.

4. Subspace Enhancement in the Modulation Domain

4.1. Introduction

Time-domain speech enhancement algorithms that are based on a subspace technique were introduced in Section 2.3. In these algorithms, the space of noisy signal vectors is decomposed into a *signal subspace* containing both speech and noise and a *noise subspace* containing only noise. The decomposition is achieved by the Karhunen-Loéve Transform (KLT), an invertible linear transform that can be used to project the noisy signal vectors into a lower dimensional subspace that preserves almost all the signal energy [45]. The key assumption underlying this approach is that the covariance matrix of the clean speech vector is close to rank-deficient. The validity of the assumption is a consequence of the Linear Predictive Coding (LPC) model of speech production in which a speech signal is generated by a low-order autogressive process. In this chapter it will be shown that it is possible to apply a subspace enhancement approach successfully to the modulation domain rather than the time domain. As was shown in Chapter 3, the speech spectral amplitude envelope of each frequency bin can be well represented by a low-order LPC model, and the modulation domain algorithms in [66, 75] implicitly make this as-

sumption. The strong temporal correlation of the sequence of spectral amplitudes with a frequency bin means that the vector of the spectral amplitudes may also be decomposed into a signal subspace and a noise space. To confirm the validity of this, the eigenvalues of the covariance matrix of the modulation domain speech vector $\mathbf{s}_l = \left[\begin{array}{ccc} S_l(0,k) & \cdots & S_l(L-1,k) \end{array} \right]^T$ are examined, where $S_l(n,k)$ is defined in Section 1.2.3. It is shown in Figure 4.1 the ordered eigenvalues of the covariance matrix of modulation-domain speech vector, $\mathbf{R}_S = \mathrm{E}\left(\mathbf{s}_l \mathbf{s}_l^T\right)$, averaged over the entire TIMIT core test set using the framing parameters defined in Section 4.4.1 with a modulation frame length $L = 32$, where $\mathrm{E}(\cdot)$ denotes the expected value. It can be seen that the eigenvalues decrease rapidly and that most of the speech energy is included in the first 10 eigenvalues. Based on this observation, this chapter will extend the subspace enhancement approach to the modulation domain.

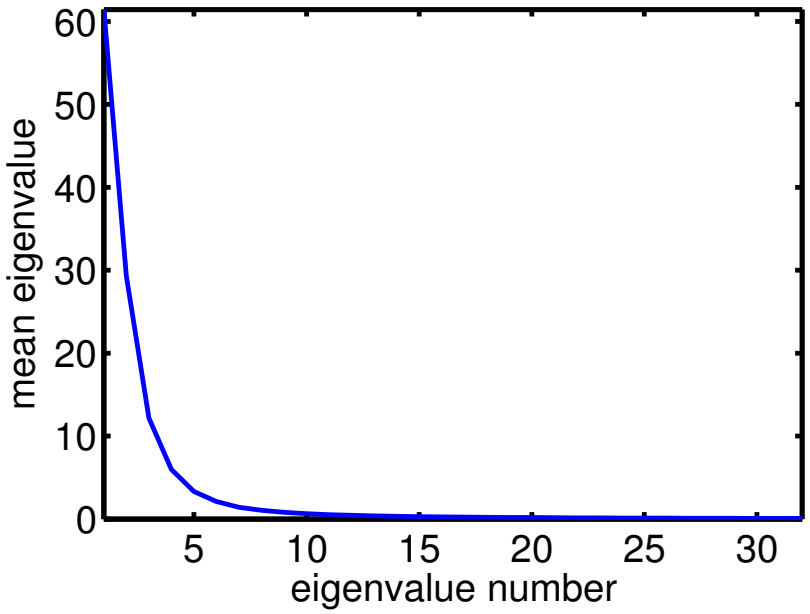

Figure 4.1.: Mean eigenvalues of covariance matrix of clean speech from the TIMIT database.

4.2. Subspace method in the short-time modulation domain

The block diagram of the proposed modulation-domain subspace enhancer is shown in Figure 4.2. The noisy speech $z(t)$ is first transformed into the acoustic domain using a STFT to obtain a sequence of spectral envelopes $|Z_{n,k}|e^{j\theta_{n,k}}$ where $|Z_{n,k}|$ is the spectral amplitude of frequency bin k in frame n. The sequence $|Z_{n,k}|$ is now divided into overlapping windowed modulation frames of length L with a frame increment Q giving $Z_l(n,k) = \check{h}_n|Z_{lQ+n,k}|$ for $n = 0, \cdots, L-1$ where $\check{h}_n$ is a modulation-domain window function. A Time Domain Constraint (TDC) subspace technique, which is described in Section 2.3, is applied independently to each frequency bin within each modulation frame to obtain the estimated clean speech spectral amplitudes $\widehat{S}_l(n,k)$ in frame l. The reason why the TDC estimator rather than Spectral Domain Constraint (SDC) estimator is chosen for the enhancer is that it has been shown in [46] that, for colored noise, the TDC estimator performs better than SDC estimator and, as noted in Section 2.5, any type of noise in the time domain is colored in the modulation domain because of the correlation introduced by the overlap between the acoustic frames.

After the modulation domain speech vector is estimated by the TDC estimator, the modulation frames are combined using overlap-addition to obtain the estimated clean speech envelope sequence $|\widehat{S}_{n,k}|$ and these are then combined with the noisy speech phases $\theta_{n,k}$ and an ISTFT is applied to give the estimated clean speech signal $\hat{s}(t)$.

As with the modulation domain Kalman filter described in Section 2.5, a linear

model in the spectral amplitude domain is assumed

$$Z_l(n,k) = S_l(n,k) + W_l(n,k) \tag{4.1}$$

where S and W denote the modulation frames of clean speech and noise respectively. Since each frequency bin is processed independently, the frequency index, k, will be omitted in the remainder of this section. The modulation domain speech vector, $\mathbf{s}_l$, has been defined in Section 4.1. In an analogous way, the noisy speech vector, $\mathbf{z}_l$, and noise vector, $\mathbf{w}_l$, are defined. If $\mathbf{R}_Z$ and $\mathbf{R}_W$ are defined similarly to $\mathbf{R}_S$, and because the spectral amplitudes of the speech and noise are assumed to be additive in (4.1), the covariance matrices of the speech and noise are also additive, which is

$$\mathbf{R}_Z = \mathbf{R}_S + \mathbf{R}_W$$

Thus, if $\mathbf{R}_W$ is known, the eigen-decomposition can be performed

$$\mathbf{R}_W^{-\frac{1}{2}}\mathbf{R}_Z\mathbf{R}_W^{-\frac{1}{2}} = \mathbf{R}_W^{-\frac{1}{2}}\mathbf{R}_S\mathbf{R}_W^{-\frac{1}{2}} + \mathbf{I} = \mathbf{U}\mathbf{P}\mathbf{U}^T \tag{4.2}$$

where $\mathbf{R}_W^{\frac{1}{2}}$ is the positive definite square root of $\mathbf{R}_W$. From this the whitened clean speech eigenvalues can be estimated as

$$\Lambda = \max(\mathbf{P} - \mathbf{I}, 0) \tag{4.3}$$

the operator $\max(\cdot)$ is placed to prevent the eigenvalues becoming negative which may otherwise happen due to errors in the estimate of $\mathbf{P}$. The clean speech vector from the noisy vector using a linear estimator, $\mathbf{H}_l$, will be estimated as

$$\hat{\mathbf{s}}_l = \mathbf{H}_l\mathbf{z}_l \tag{4.4}$$

It has been shown in [48] that the optimal TDC linear estimator is given by

$$\mathbf{H}_l = \mathbf{R}_W^{\frac{1}{2}} \mathbf{U} \Lambda (\Lambda + \eta \mathbf{I})^{-1} \mathbf{U}^T \mathbf{R}_W^{-\frac{1}{2}} \tag{4.5}$$

where η controls the tradeoff between speech distortion and noise suppression.

The estimator in (4.5) has been given in (2.6) and a detailed derivation of (4.5) has been given in Section 2.3. The action of the estimator in (4.5) can be interpreted as first whitening the noise with $\mathbf{R}_W^{-\frac{1}{2}}$ and then applying a KLT, $\mathbf{U}^T$, to perform the subspace decomposition. In the transform domain, the gain matrix, $\Lambda(\Lambda + \eta \mathbf{I})^{-1}$, projects the vector into the signal subspace and attenuates the noise components by a factor controlled by η, discussed in Section 4.4.1 for the time-domain enhancer.

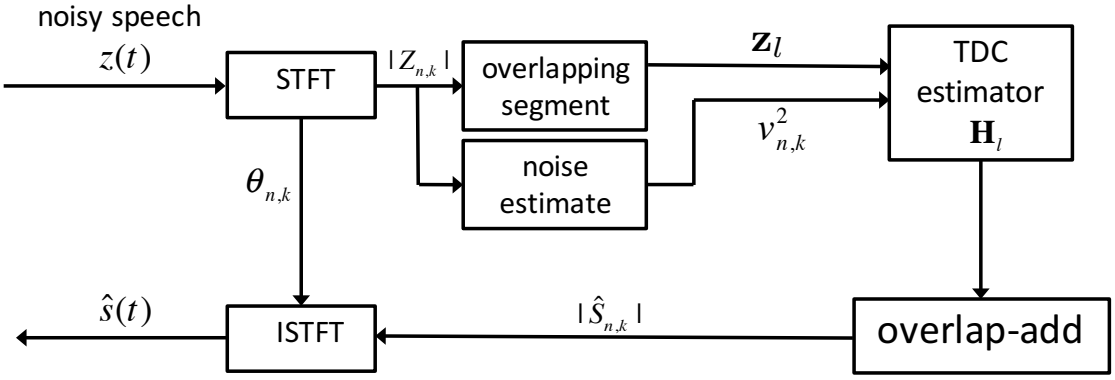

Figure 4.2.: Diagram of proposed short-time modulation domain subspace enhancer.

4.3. Noise Covariance Matrix Estimation

Now the estimation of the noise covariance matrix $\mathbf{R}_W(k)$ is considered. For quasi-stationary noise, $\mathbf{R}_W(k)$ will be a symmetric Toeplitz matrix whose first column is given by the autocorrelation vector $\mathbf{a}_c(k) = \begin{bmatrix} a_c(0,k) & \cdots & a_c(L-1,k) \end{bmatrix}^T$ where

$a_c(\tau,k) = \mathrm{E}\left(|W_{n,k}|\,|W_{n+\tau,k}|\right)$. This section will begin by determining $a_c(\tau,k)$ for the case when $w(t)$ is white noise and then extend this to colored noise.

First suppose $w(t) \sim \mathcal{N}(0,\sigma_w^2)$ is a zero-mean Gaussian white noise signal. If the acoustic frame length is T samples with a frame increment of M samples, the output of the initial STFT stage in Figure 4.2 is

$$\widetilde{W}_{n,k} = \sum_{t=0}^{T-1} w(nM+t)h(t)e^{-2\pi j\frac{tk}{T}} \tag{4.6}$$

where $h(t)$ is the acoustic window function and the complex spectral coefficients, $\widetilde{W}_{n,k}$, have a zero-mean complex Gaussian distribution [7]. The expectation, $\mathrm{E}\left(\widetilde{W}_{n,k}\widetilde{W}^*_{n+\tau,k}\right)$, where $*$ denotes complex conjugation, is given by

$$\begin{aligned}
&\mathrm{E}\left(\widetilde{W}_{n,k}\widetilde{W}^*_{n+\tau,k}\right)\\
&= \mathrm{E}\left(\sum_{t,s=0}^{T-1} w(nM+t)h(t)w(nM+s+\tau M)h(s)e^{-2\pi j\frac{(t-s)k}{T}}\right)\\
&= \sigma_w^2\sum_{t=0}^{T-1} h(t)h(t-\tau M)e^{-2\pi j\frac{\tau Mk}{R}}
\end{aligned} \tag{4.7}$$

since, for white noise,

$$\mathrm{E}\left(w(nM+t)w(nM+s+\tau M)\right) = \sigma_w^2\delta\left(t-s-\tau M\right).$$

By setting $\tau = 0$, therefore the spectral power of the white noise in any frequency bin can be obtained as

$$\nu_w^2 = \mathrm{E}\left(\left|\widetilde{W}_{n,k}\right|^2\right) = \sigma_w^2\sum_{t=0}^{T-1} h^2(t). \tag{4.8}$$

Defining

$$\rho_h(\tau,k) = \frac{\sum_{t=0}^{T-1} h(t)h(t-\tau M)e^{-2\pi j\frac{\tau Mk}{T}}}{\sum_{t=0}^{T-1} h^2(t)}$$

now (4.7) and (4.8) can be used to write

$$\mathrm{E}\left(\widetilde{W}_{n,k}\widetilde{W}^{*}_{n+\tau,k}\right) = \nu_w^2 \rho_h(\tau,k) \tag{4.9}$$

where $\rho_h(\tau,k)$ depends on the window, $h(t)$, but not on the noise variance ν_w^2.

Now the autocorrelation sequence of the short-time Fourier coefficients, $\mathrm{E}\left(\widetilde{W}_{n,k}\widetilde{W}^{*}_{n+\tau,k}\right)$, has been obtained. From [88, pp. 95-97] the autocorrelation sequence of their magnitudes can be further obtained as

$$\begin{aligned} a_c(\tau,k) &= \mathrm{E}\left(\left|\widetilde{W}_{n,k}\right|\left|\widetilde{W}_{n+\tau,k}\right|\right) \\ &= \frac{\pi}{4}\nu_w^2 \times {}_2F_1\left(-\frac{1}{2},-\frac{1}{2},1;\left|\rho_h(\tau,k)\right|^2\right) \end{aligned} \tag{4.10}$$

where ${}_2F_1(\cdots)$ is the Gauss hypergeometric function [89], the definition of which is given in Section A.1.1 of Appendix A. The details of the derivation of 4.10 are given in Section B.2 of Appendix B.

Therefore, if define

$$\mathbf{a}_0(k) = \nu_w^{-2}\left[\begin{array}{ccc} a_c(0,k) & \cdots & a_c(L-1,k) \end{array}\right]^T$$

and $\mathbf{R}_0(k)$ is a symmetric Toeplitz matrix with $\mathbf{a}_0(k)$ as the first column, the noise covariance matrix can be obtained as

$$\mathbf{R}_W(k) = \nu_w^2 \mathbf{R}_0(k) \tag{4.11}$$

where $\mathbf{R}_0(k)$ does not depend on ν_w^2.

Assuming that $w(t)$ is quasi-stationary colored noise with a correlation time that is small compared with the acoustic frame length, $\widetilde{W}_{n+\tau,k}$ will be multiplied by a factor that depends on the frequency index, k, but not on τ [90]. In this case, the

previous analysis still applies but, for frame l, (4.11) now becomes

$$\mathbf{R}_W(k) = \nu_l^2(k)\mathbf{R}_0(k) \tag{4.12}$$

where $\nu_l^2(k) = \mathrm{E}\left(|W(lQ,k)|^2\right)$ is the noise power spectrum corresponding to the modulation frame l and, as shown above, $\mathbf{R}_0(k)$ is independent of the noise power spectrum. This means that $\mathbf{R}_W(k)$ can be estimated directly from an estimate of $\nu_l^2(k)$ which can be obtained from the noisy speech signal, $y(t)$, using a noise power spectrum estimator such as [41] or [43].

Substituting (4.12) into (4.2)-(4.5), the following equations can be obtained

$$\begin{aligned}
\mathbf{R}_0^{-\frac{1}{2}}(k)\mathbf{R}_Z(k)\mathbf{R}_0^{-\frac{1}{2}}(k) &= \mathbf{U}(k)\overline{\mathbf{P}}(k)\mathbf{U}^T(k) \\
\overline{\Lambda}(k) &= \max(\overline{\mathbf{P}}(k) - \nu_l^2(k)\mathbf{I}, 0) \\
\mathbf{H}_l(k) &= \mathbf{R}_0^{\frac{1}{2}}(k)\mathbf{U}(k)\overline{\Lambda}(k)(\overline{\Lambda}(k) + \eta\nu_l^2(k)\mathbf{I})^{-1}\mathbf{U}(k)^T\mathbf{R}_0^{-\frac{1}{2}}(k)
\end{aligned}$$

in which the whitening transformation, $\mathbf{R}_0^{-\frac{1}{2}}(k)$, can be precomputed since it depends only on the window, $h(t)$, and is independent of the noise power spectrum. In addition, because the matrix $(\overline{\Lambda}(k) + \eta\nu_l^2(k)\mathbf{I})$ is a diagonal matrix whose inverse is straightforward to calculate, the computational complexity of the estimator is greatly reduced.

To confirm the validity of the analysis given above, the autocorrelation vector, $\mathbf{a}_c$, has been evaluated for the 'F16' noise in the RSG-10 database [25] using the framing parameters given in Section 4.4.1 with a modulation frame length $L = 32$. Figure 4.3 shows the true autocorrelation averaged over all k together with the autocorrelation from (4.10) using the true noise periodogram. It can be seen that the two curves match very closely and that for $\tau \geq \frac{R}{J} = 4$, the STFT analysis windows do not

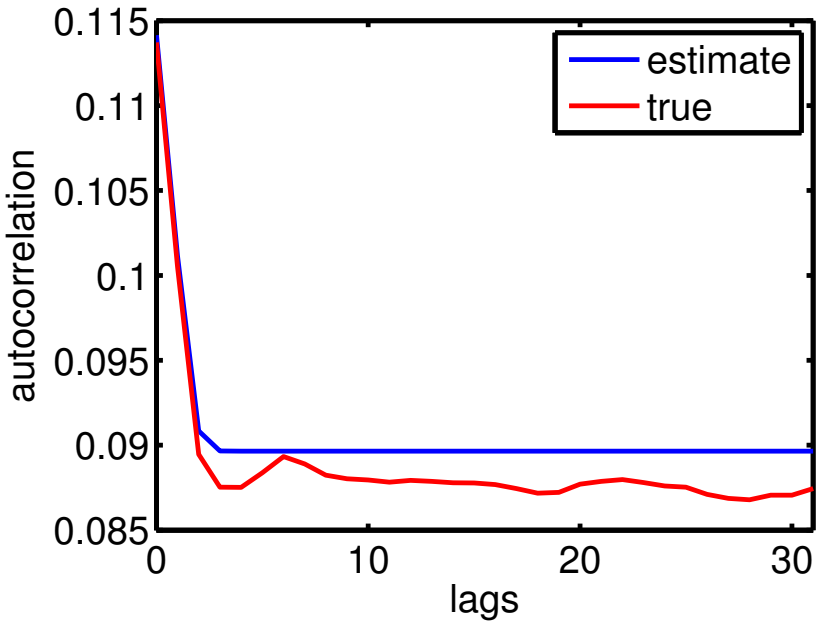

Figure 4.3.: Estimated and true value of the average autocorrelation sequence in one modulation frame.

overlap and so $a(\tau, k)$ is constant.

4.4. Evaluation and Conclusions

4.4.1. Implementation and experimental results

In this section, the proposed Modulation Domain Subspace (MDSS) enhancer is compared with the TDC version of the Time Domain Subspace (TDSS) enhancer from [46] and the Modulation Domain Spectral Subtraction (MDST) enhancer from [65] using the default parameters. Compared to the proposed MDSS enhancer, TDSS enhancer applies the subspace method in the time domain instead of the modulation domain, while MDST enhancer applies the spectral subtraction method rather than subspace method in the modulation domain. In our experiments, the core test set from the TIMIT database is used and the speech is corrupted by 'white', 'factory2' and 'babble' noise from [25] at -5, 0, 5, 10, 15 and 20 dB SNR

(see Chapter 2 for more details). The algorithm parameters were determined by optimizing performance on the development set described in Section 1.4.1.1 and the parameters are listed in Table 4.1.

Parameter	Settings
Sampling frequency	8 kHz
Acoustic frame length	16 ms
Acoustic frame increment	4 ms
Modulation frame length	128 ms
Modulation frame increment	16 ms
Analysis-synthesis window	Hamming window

Table 4.1.: Parameter settings in experiments.

Additionally, the noise power spectrum was estimated using the algorithm in [43, 85] and, following [46], the factor η in (4.5) was selected as

$$\eta = \begin{cases} 5 & \mathrm{SNR}_{\mathrm{dB}} \leq -5 \\ \eta_0 - (\mathrm{SNR}_{\mathrm{dB}})/6.25 & -5 < \mathrm{SNR}_{\mathrm{dB}} < 20 \\ 1 & \mathrm{SNR}_{\mathrm{dB}} \geq 20 \end{cases}$$

where $\eta_0 = 4.2$, $\mathrm{SNR}_{\mathrm{dB}} = 10\log_{10}(\mathrm{tr}(\Lambda)/L)$ and the operator $\mathrm{tr}(\cdot)$ calculates the trace of the diagonal matrix Λ.

To avoid any of the estimated spectral amplitudes in $\hat{\mathbf{s}}_l$ becoming negative, a floor equal to 20 dB below the corresponding noisy spectral amplitudes in $\mathbf{z}_l$ is set, so that (4.4) now becomes

$$\hat{\mathbf{s}}_l = \max(\mathbf{H}_l \mathbf{z}_l, 0.1\mathbf{z}_l) \tag{4.13}$$

The performance of the three speech enhancers are evaluated and compared using the segmental SNR (segSNR) and Perceptual Evaluation of Speech Quality (PESQ) measure, averaged over all the sentences in the core TIMIT test set. The average segSNR for the noisy speech, TDSS enhancer [46], MDST enhancer [65] and the

proposed MDSS enhancer is shown for factory noise in Figure 4.4 as a function of the global SNR of the noisy speech. It can be seen that MDSS enhancer and TDSS enhancer give similar performance at low SNRs. At SNRs higher than 10 dB, MDSS enhancer performs better than TDSS enhancer, giving segSNR improvement of about 3 dB at 20 dB SNR. The equivalent figures for babble noise and white noise are given in Figure 4.5 and Figure 4.6, respectively. For babble noise, it shows a same trend in performance as that of the factory noise and at low SNRs, MDSS enhancer also performs slight better than TDSS enhancer. For white noise, MDSS enhancer gives a better performance than TDSS enhancer at SNRs higher than 15 dB and at 20 dB, it gives segSNR improvement of about 1.5 dB. However, at lower SNRs, TDSS shows a better performance and at -5 dB it gives a improvement of about 2 dB over the MDSS algorithm. The corresponding PESQ plots are shown in Figures 4.8 to 4.9, for noisy speech corrupted by factory noise, babble noise and white noise respectively at different global SNRs, and the corresponding enhanced speech by the three enhancers mentioned above.

It can be seen that, as the results implied by the segSNR, for colored noise, the proposed MDSS enhancer performs better than the other two enhancers, especially at low SNRs which gives a PESQ improvement of more than 0.2 over a wide range of SNRs. For white noise, however, the performance of the MDSS enhancer is not as good as the TDSS enhancer, except at very low SNRs. In order to understand why the TDSS algorithm is better for white noise than MDSS enhancer, the performance of the TDSS and MDSS algorithms for speech-shaped noise is explored. The speech-shaped noise is a random noise that has the same long-term spectrum as a given speech signal, which is a stationary colored noise. The segSNR and PESQ of the three algorithms are given in Figs. 4.10 and 4.11, respectively. It can be seen that, although the segSNR of the TDSS enhancer is better than that resulting

from the MDSS enhancer, the MDSS gives better performance in PESQ over the TDSS enhancer, which is about 0.25 at low SNRs. By listening to the enhanced speech utterance, it can be found that the although the TDSS enhancer can reduce more background noise, it also introduce speech distortion making the speech more perceptually uncomfortable than the speech enhanced by MDSS enhancer. This finding is consistent with the results shown by the segSNR and PESQ. Based on the performance of the algorithms for the different noises, it can be seen that the MDSS algorithm, which makes use of the noise covariance estimation derived in Section 4.3, performs best for colored noise regardless of whether the noise is stationary (speech-shaped noise) or non-stationary (factory noise and babble noise). For white noise, however, the performance of the MDSS algorithm is not as good as that of the TDSS algorithms. This is not surprising because the time-domain whiteness satisfies the assumptions made in the development of the TDSS algorithms and there is no extra approximation.

Comparing the performance of the MDSS enhancer with the postprocessors proposed in Chapter 3, it can be seen that the MDSS enhancer gives similar performance for non-stationary colored noise and slightly worse performance for white noise.

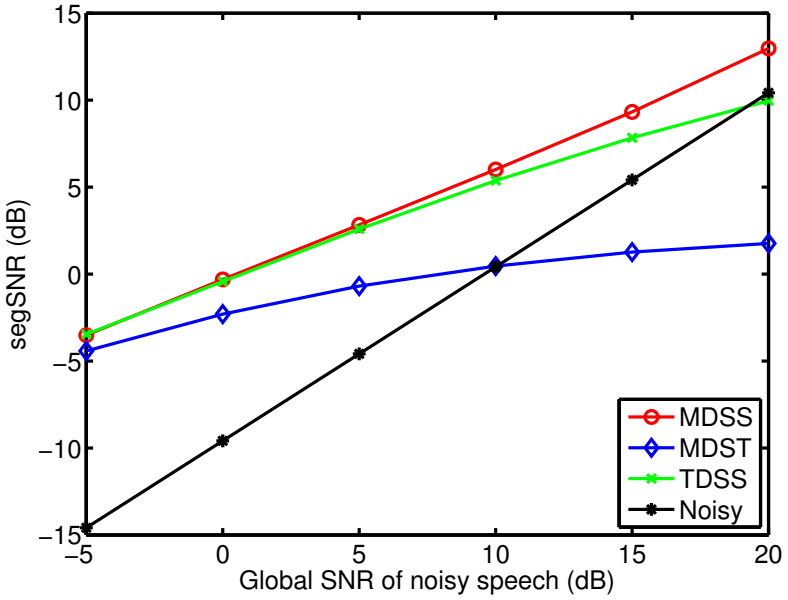

Figure 4.4.: Average segSNR values comparing different algorithms, where speech signals are corrupted by factory noise at different SNR levels. (MDSS:proposed modulation domain subspace enhancer; MDST: modulation domain spectral subtraction enhancer; TDSS: time domain subspace enhancer)

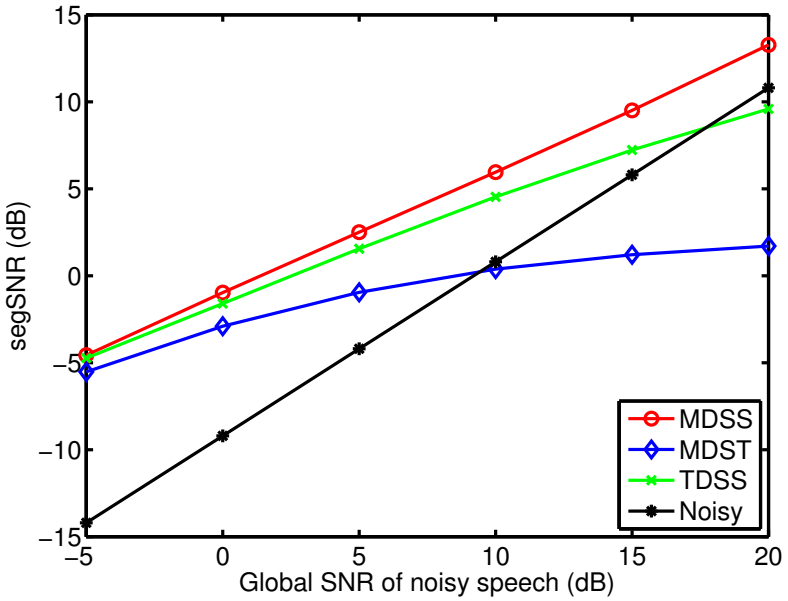

Figure 4.5.: Average segSNR values comparing different algorithms, where speech signals are corrupted by babble noise at different SNR levels.

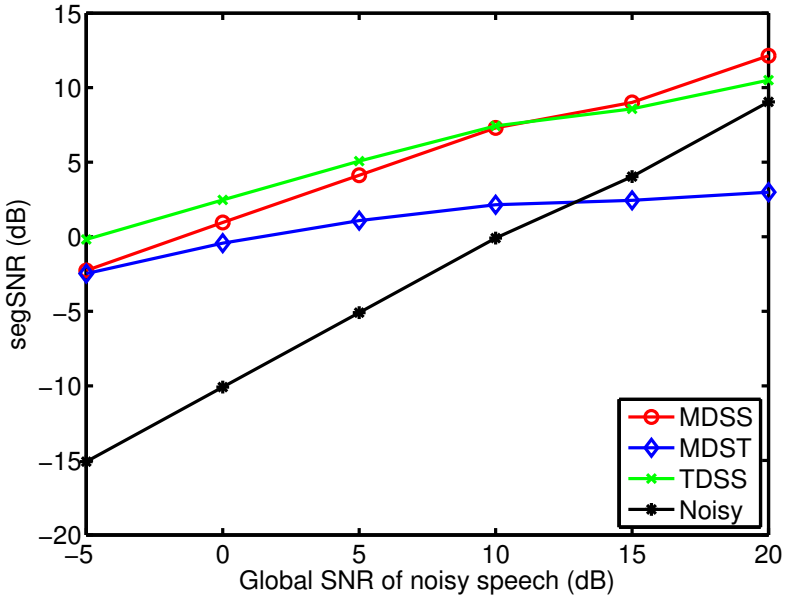

Figure 4.6.: Average segSNR values comparing different algorithms, where speech signals are corrupted by white noise at different SNR levels.

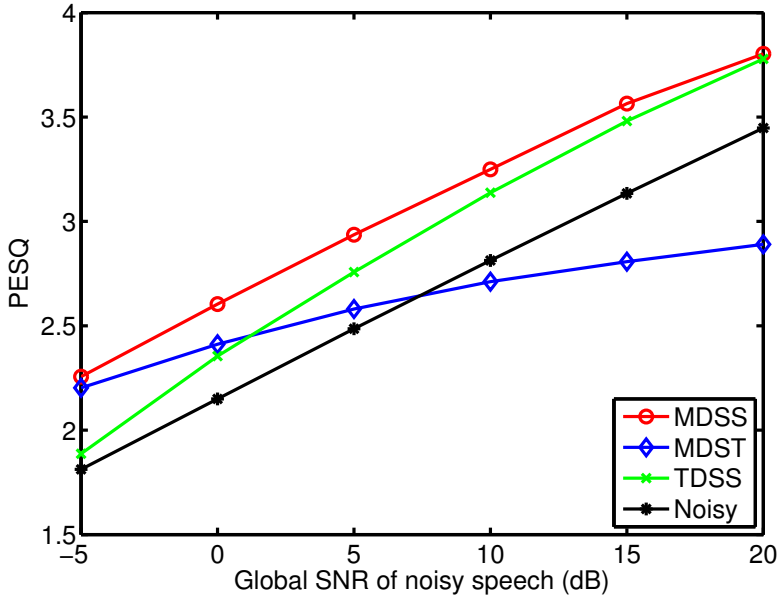

Figure 4.7.: Average PESQ values comparing different algorithms, where speech signals are corrupted by factory noise at different SNR levels.

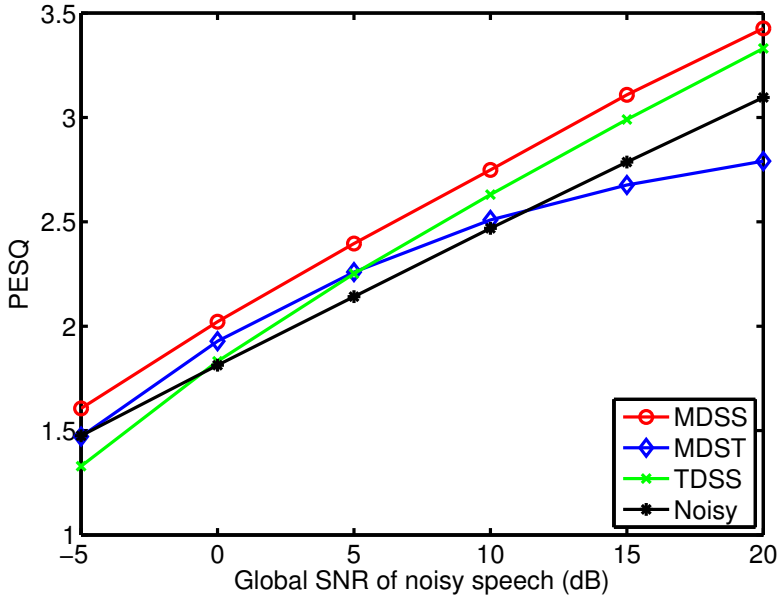

Figure 4.8.: Average PESQ values comparing different algorithms, where speech signals are corrupted by babble noise at different SNR levels.

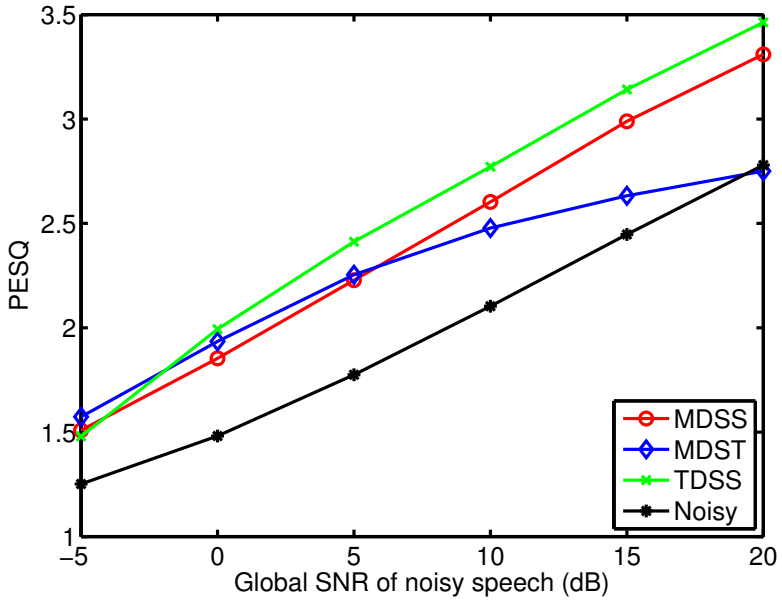

Figure 4.9.: Average PESQ values comparing different algorithms, where speech signals are corrupted by white noise at different SNR levels.

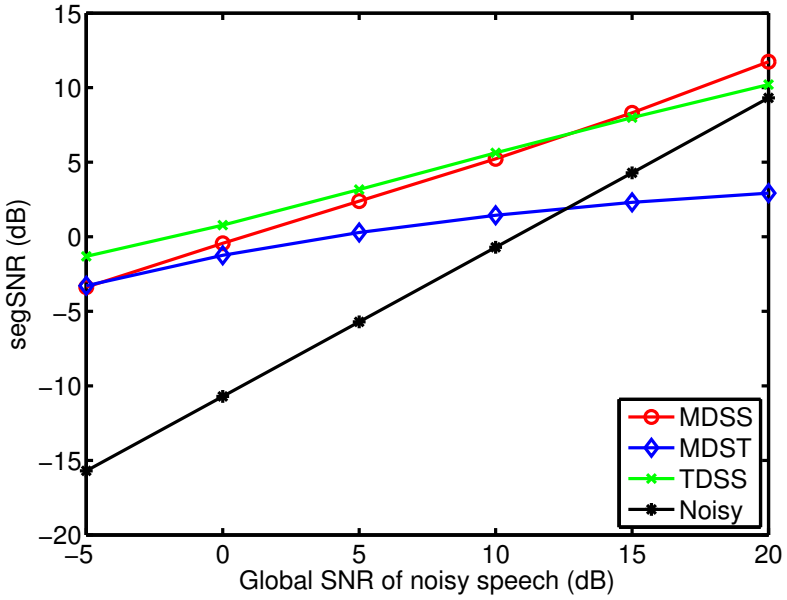

Figure 4.10.: Average segSNR values comparing different algorithms, where speech signals are corrupted by speech-shaped noise at different SNR levels.

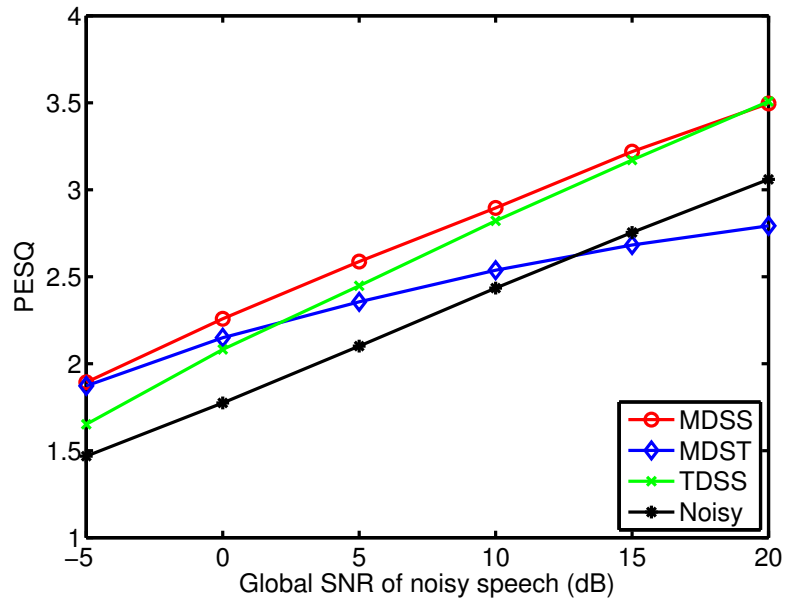

Figure 4.11.: Average PESQ values comparing different algorithms, where speech signals are corrupted by speech-shaped noise at different SNR levels.

4.4.2. Conclusions

In this chapter a speech enhancement algorithm using a subspace decomposition technique in the short-time modulation domain has been presented. It has been shown that one consequence of processing the speech in the modulation domain is that the covariance matrix is independent of the noise spectrum to within a scale factor; this means that the whitening matrix can be precomputed. The performance of the proposed enhancer has been evaluated using segSNR and PESQ and it has been shown that, for both stationary and non-stationary colored noise, it outperforms a time-domain subspace enhancer and a modulation-domain spectral-subtraction enhancer.

5. Model-based Speech Enhancement in the Modulation Domain

5.1. Overview

An overview of conventional model-based enhancement in the Short Time Fourier Transform (STFT) domain was given in Section 2.4. In this chapter, parametric models are assumed for the complex STFT coefficients of the speech and noise. The time-frequency gain function is then selected to optimize a chosen performance measure. In [7] and [53], the speech and noise STFT coefficients are both assumed to follow zero-mean complex Gaussian distributions. The noise variance is assumed to be known in advance and the ratio of the speech and noise variances, the prior SNR, is estimated recursively using the "decision-directed" approach. The two methods differ in minimizing the mean squared estimation error of either the spectral amplitude or the log spectral amplitude. A number of authors have extended the work in [7, 53] by using super-gaussian distributions for the speech amplitude prior distributions [55, 91, 54]. However the authors found that although the use of a super-gaussian prior reduced the noise level, it often did so at the expense of in-

creased speech distortion. Although these STFT-domain enhancement algorithms are able to improve the SNR dramatically, the temporal dynamics of the speech spectral amplitudes are not incorporated into the derivation of the estimator. In this chapter, two algorithms, based on the modulation domain Kalman filter, will be introduced, which combine the estimated dynamics of the spectral amplitudes with the observed noisy speech to obtain an Minimum Mean Squared Error (MMSE) estimate of the amplitude spectrum of the clean speech. Both algorithms assume that the speech and noise are additive in the complex STFT domain. The difference between the two algorithms is that the algorithm introduced in Section 5.2 only models the spectral dynamics of the clean speech while the second algorithm, presented in Section 5.3, jointly models the spectral dynamics of both speech and noise. In this chapter, a tilde diacritic, $\sim$, will be used to denote quantities relating to the estimated speech signal and a breve diacritic, $\breve{}$, will be used to denote quantities relating to the estimated noise signal.

5.2. Enhancement with Generalized Gamma prior

In this section, an MMSE spectral amplitude estimator is proposed under the assumption that the speech spectral amplitudes follow a generalized Gamma distribution [56]. The advantages of the proposed estimator over previously proposed spectral amplitude estimators [7, 56, 54] are, first, that it incorporates temporal continuity into the MMSE estimator by the use of the Kalman filter, second, that it uses a Gamma prior which is a more appropriate model for the speech spectral amplitudes than a Gaussian prior that is used in Section (2.5.1) [66] and, third, that the speech and noise are assumed to be additive in the complex STFT domain rather than in the spectral amplitude domain.

For frequency bin k of frame n, it is assumed that

$$Z_{n,k} = S_{n,k} + W_{n,k} \tag{5.1}$$

It can be seen that this assumption is different from that given in (2.15). Since each frequency bin is processed independently within our algorithm, the frequency index, k, will be omitted in the remainder of this chapter. The random variables representing the spectral amplitudes are denoted as: $A_n = |S_n|$, $R_n = |Z_n|$ and $F_n = |W_n|$. The prediction model assumed for the clean speech spectral amplitude is the same as that defined in Section 2.5.1, which is given by

$$\tilde{\mathbf{s}}_n = \widetilde{\mathbf{A}}_n \tilde{\mathbf{s}}_{n-1} + \tilde{\mathbf{d}} e_n \tag{5.2}$$

where $\tilde{\mathbf{s}}_n$ denotes the state vector of speech amplitudes and $\widetilde{\mathbf{A}}_n$ denotes the transition matrix for the speech amplitudes. $\tilde{\mathbf{d}} = [1\ 0\ \cdots\ 0]^T$ is a p-dimensional vector and the speech transition matrix has the form

$$\widetilde{\mathbf{A}}_n = \begin{bmatrix} -\tilde{\mathbf{b}}_n^T \\ \mathbf{I} \quad \mathbf{0} \end{bmatrix} \tag{5.3}$$

where $\tilde{\mathbf{b}} = [b_1\ \cdots\ b_p]^T$ is the LPC coefficient vector, and $\mathbf{0}$ denotes an all-zero column vector of length $p-1$. The prediction residual signal, $\tilde{e}_n$, is assumed to have zero mean and variance $\tilde{\sigma}^2$.

5.2.1. Proposed estimator description

A block diagram of the proposed algorithm is shown in Figure 5.1. The noise estimator block uses the noisy speech amplitudes, $R_{n,k}$, to estimate the prior noise

power spectrum, $\nu^2_{n,k}$, in each frame using one of the noise estimation algorithms which are introduced in Section 2.2, such as [92] and [43]; this noise estimate is then sent both to the Kalman Filter block and also to a conventional log-amplitude MMSE (logMMSE) enhancer [53]. Within the "Modulation Domain LPC" block, the enhanced speech from the logMMSE enhancer is divided into overlapping modulation frames and Linear Predictive Coding (LPC) analysis is performed separately in each frequency bin, k. Autocorrelation LPC [22] is performed on each modulation frame to determine the coefficients, $\widetilde{\mathbf{b}}_n$, and thence the transition matrix $\widetilde{\mathbf{A}}_n$ defined in (5.3).

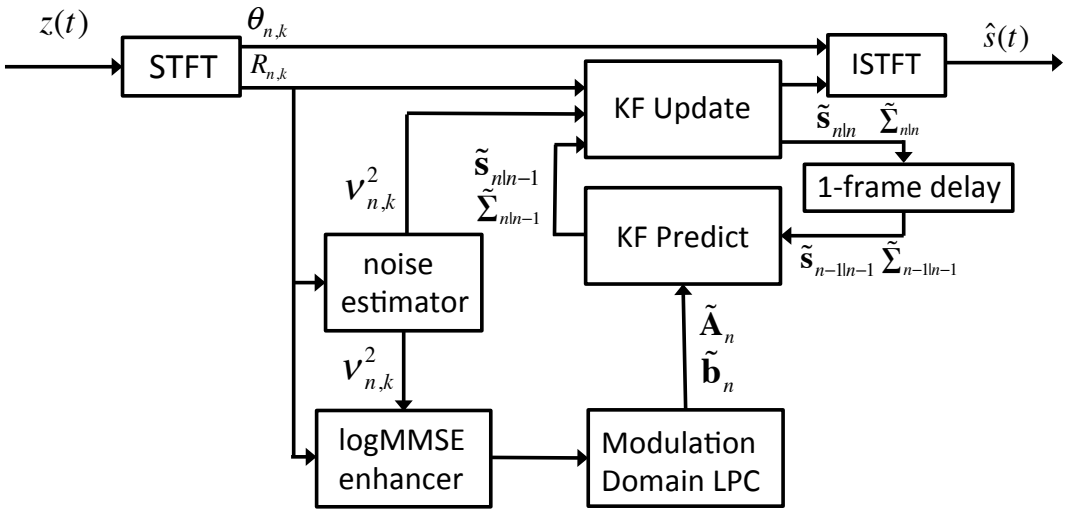

Figure 5.1.: Diagram of KFMMSE algorithm

5.2.2. Kalman filter prediction step

The Kalman filter prediction step ("KF Predict" in Figure 5.1) estimates the state vector mean and covariance at time n, $\tilde{\mathbf{s}}_{n|n-1}$ and $\tilde{\mathbf{\Sigma}}_{n|n-1}$, from their values at time $n-1$, $\tilde{\mathbf{s}}_{n-1|n-1}$ and $\tilde{\mathbf{\Sigma}}_{n-1|n-1}$.

First, the time update model equations are rewritten:

$$\tilde{\mathbf{s}}_{n|n-1} = \tilde{\mathbf{A}}_n \tilde{\mathbf{s}}_{n-1|n-1} \tag{5.4}$$

$$\tilde{\mathbf{\Sigma}}_{n|n-1} = \tilde{\mathbf{A}}_n \tilde{\mathbf{\Sigma}}_{n-1|n-1} \tilde{\mathbf{A}}_n^T + \widetilde{\mathbf{Q}}_n \tag{5.5}$$

where $\widetilde{\mathbf{Q}}_n = e_n \tilde{\mathbf{d}} \tilde{\mathbf{d}}^T$. The first element of the state vector, $\tilde{\mathbf{s}}_{n|n-1}$, corresponds to the spectral amplitude in the current frame, $A_{n|n-1}$, and so its prior mean and variance are given by

$$\tilde{\mu}_{n|n-1} \triangleq E(A_n|\mathcal{R}_{n-1}) = \tilde{\mathbf{d}}^T \tilde{\mathbf{s}}_{n|n-1} \tag{5.6}$$

$$\tilde{\sigma}^2_{n|n-1} \triangleq Var(A_n|\mathcal{R}_{n-1}) = \tilde{\mathbf{d}}^T \tilde{\mathbf{\Sigma}}_{n|n-1} \tilde{\mathbf{d}}, \tag{5.7}$$

where $\mathcal{R}_n = [R_1 \dots R_n]$ represents the observed speech amplitudes up to time n and $\tilde{\mathbf{d}} = [1\, 0 \dots 0]^T$.

5.2.3. Kalman filter MMSE update model

In this section, the Kalman filter MMSE update step ("KF Update" in Figure 5.1) is described which determines an updated state estimate by combining the predicted state vector and covariance, the estimated noise and the observed spectral amplitude. Within the update step, the distribution of the prior speech amplitude $A_{n|n-1}$ is modeled using a 2-parameter Gamma distribution

$$p\left(a_n|\mathcal{R}_{n-1}\right) = \frac{2a_n^{2\gamma_n - 1}}{\beta_n^{2\gamma_n}\Gamma\left(\gamma_n\right)} \exp\left(-\frac{a_n^2}{\beta_n^2}\right), \tag{5.8}$$

where $\Gamma(\cdot)$ is the Gamma function. The distribution is obtained by setting $d = 2$ in the generalized Gamma distribution given in Section 2.13 in Chapter 2, and the

two parameters, β_n and γ_n are chosen to match the mean μ_n and variance σ_n^2 of the predicted amplitude from (5.6) and (5.7). Examples of the probability density functions from (5.8) with variance, $\sigma^2 = 1$ and means, μ, in the range 0.5 to 8 are shown in Figure 5.2, from which it can be seen that the distribution in (5.8) is sufficiently flexible to model the outcome of the prediction over a wide range of μ_n/σ_n. It worth nothing that the prior knowledge about A_n depends on the observed speech amplitudes up to time $n-1$, $\mathcal{R}_{n-1}$, rather than on the estimate of the speech amplitude at time $n-1$, A_{n-1}.

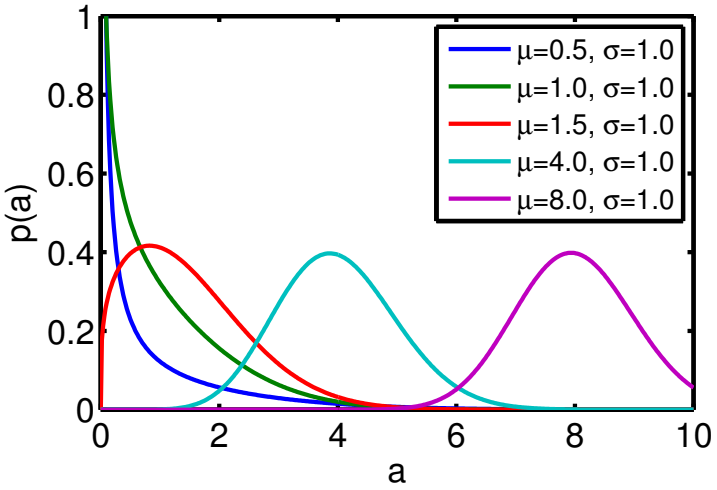

Figure 5.2.: Curves of Gamma probability density function for (5.8) with variance $\sigma^2 = 1$ and different means.

At frame n, the mean and variance of the Gamma distribution in (5.8) can be expressed in terms of β_n and γ_n [93] as

$$\tilde{\mu}_{n|n-1} = \beta_n \frac{\Gamma(\gamma_n + 0.5)}{\Gamma(\gamma_n)}, \tag{5.9}$$

$$\tilde{\sigma}^2_{n|n-1} = \beta_n^2 \left(\gamma_n - \frac{\Gamma^2(\gamma_n + 0.5)}{\Gamma^2(\gamma_n)} \right). \tag{5.10}$$

β between (5.9) and (5.10) can be eliminated to obtain

$$\frac{\Gamma^2(\gamma_n + 0.5)}{\gamma_n \Gamma^2(\gamma_n)} = \frac{\tilde{\mu}^2_{n|n-1}}{\tilde{\mu}^2_{n|n-1} + \tilde{\sigma}^2_{n|n-1}} \triangleq \lambda_n \tag{5.11}$$

the non-linear equation (5.11) needs to solve to determine γ_n from the value of λ_n which can be calculated from $\tilde{\mu}_{n|n-1}$ and $\tilde{\sigma}^2_{n|n-1}$ and which will always satisfy $0 < \lambda_n < 1$. Instead of dealing with γ_n directly, it is convenient to set $\varphi_n = \arctan(\gamma_n)$ where φ_n lies in the range $0 < \varphi_n < \frac{\pi}{2}$. The solid line in Figure 5.3 shows the function $\varphi_n(\lambda_n)$. This function can be approximated well with a low-order polynomial that is constrained to pass through the points $(0, 0)$ and $(1, \frac{\pi}{2})$ and in the experiments in Section 5.2.7 the quartic approximation is used

$$\varphi_n(\lambda_n) = -0.1640\lambda_n^4 + 2.3612\lambda_n^3 - 1.2182\lambda_n^2 + 0.5918\lambda_n$$

which is shown with asterisks in Figure 5.3. Given λ_n this polynomial can be used to obtain φ_n and thence γ_n by the inverse transform $\gamma_n = \tan(\varphi_n)$.

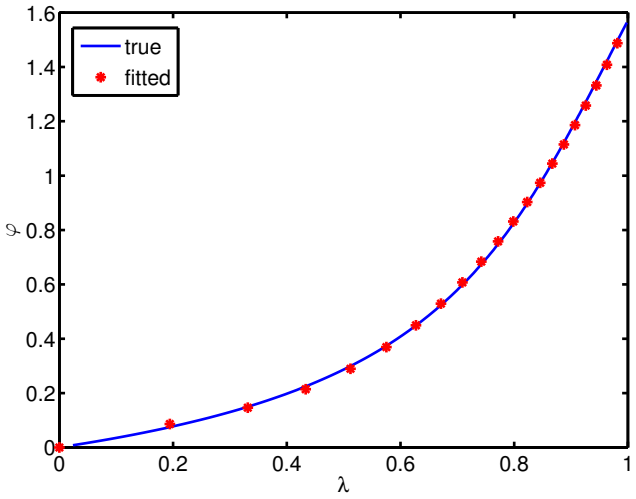

Figure 5.3.: The curve of φ versus λ, where $0 < \varphi = \arctan(\gamma) < \frac{\pi}{2}$ and $0 < \lambda = \frac{\Gamma^2(\gamma+0.5)}{\Gamma^2(\gamma)\gamma} < 1$.

5.2.4. Derivation of the estimator

The MMSE estimate of A_n is given by the conditional expectation

$$\tilde{\mu}_{n|n} = E(A_n|\mathcal{R}_n) = \int_0^\infty a_n p(a_n|\mathcal{R}_n) da_n. \tag{5.12}$$

Using Bayes rule, the conditional probability is expressed as

$$p\left(a_n|\mathcal{R}_n\right) = p(a_n|z_n, \mathcal{R}_{n-1}) = \frac{\int_0^{2\pi} p\left(z_n|a_n, \phi_n, \mathcal{R}_{n-1}\right) p\left(a_n, \phi_n|\mathcal{R}_{n-1}\right) d\phi_n}{p\left(z_n|\mathcal{R}_{n-1}\right)} \tag{5.13}$$

where ϕ_n is the realization of the random variable Φ_n which represents the phase of the clean speech. Because Z_n is conditionally independent of $\mathcal{R}_{n-1}$ given a_n and ϕ_n, (5.13) becomes

$$p\left(a_n|\mathcal{R}_n\right) = \frac{\int_0^{2\pi} p\left(z_n|a_n, \phi_n\right) p\left(a_n, \phi_n|\mathcal{R}_{n-1}\right) d\phi_n}{p\left(z_n|\mathcal{R}_{n-1}\right)}. \tag{5.14}$$

Following [7], the observation noise is assumed to be complex Gaussian distributed with variance $\nu_n^2 = E(|W_n|^2)$ leading to the observation prior model

$$p(z_n|a_n, \phi_n) = \frac{1}{\pi\nu_n^2} \exp\left\{-\frac{1}{\nu_n^2}|z_n - a_n e^{j\phi_n}|^2\right\}. \tag{5.15}$$

Under the assumption of the statistical models previously defined it is assumed that the phase components and amplitude components, Φ_n and A_n, are independent and Φ_n is uniformly distributed on the interval $[0, 2\pi]$. The posterior distribution of the

speech amplitude, $p(a_n|\mathcal{R}_n, \phi_n)$, can now be found and it is given by

$$\begin{aligned} p(a_n|\mathcal{R}_n) &= \frac{\int_0^{2\pi} p\left(z_n|a_n,\phi_n\right) p\left(a_n,\phi_n|\mathcal{R}_{n-1}\right) d\phi_n}{p\left(z_n|\mathcal{R}_{n-1}\right)} \\ &= \frac{\int_0^{2\pi} p\left(z_n|a_n,\phi_n\right) p\left(a_n,\phi_n|\mathcal{R}_{n-1}\right) d\phi_n}{\int_0^{\infty}\int_0^{2\pi} p\left(z_n|a_n,\phi_n\right) p\left(a_n,\phi_n|\mathcal{R}_{n-1}\right) d\phi_n da_n} \\ &= \frac{\int_0^{2\pi} \frac{a_n^{2\gamma_n-1}}{\pi^2\Gamma(\gamma_n)\beta_n^{2\gamma_n}\nu_n^2} \exp\left\{-\frac{a_n^2}{\beta_n^2} - \frac{1}{\nu_n^2}|z_n - a_n e^{j\phi_n}|^2\right\} d\phi_n}{\int_0^{\infty}\int_0^{2\pi} \frac{a_n^{2\gamma_n-1}}{\pi^2\Gamma(\gamma_n)\beta_n^{2\gamma_n}\nu_n^2} \exp\left\{-\frac{a_n^2}{\beta_n^2} - \frac{1}{\nu_n^2}|z_n - a_n e^{j\phi_n}|^2\right\} d\phi_n da_n}. \end{aligned} \tag{5.16}$$

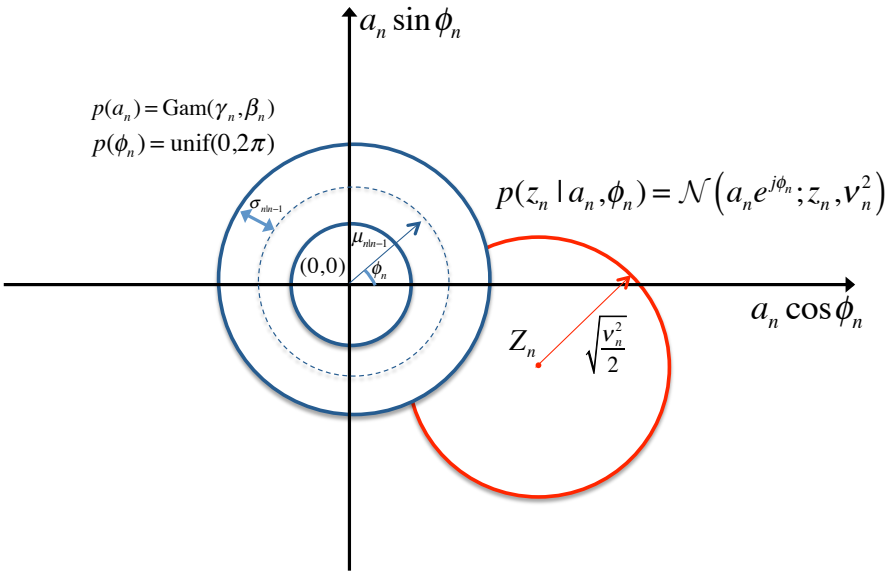

Figure 5.4.: Statistical model assumed in the derivation of the posterior estimate, where blue ring-shape distribution centered on the origin represents the prior model while the red circle centered on the observation, Z_n, represents the observation model.

The model assumed in equation (5.16) is shown in Figure 5.4, where blue ring-shape distribution centered on the origin represents the prior model, $p\left(a_n, \phi_n|\mathcal{R}_{n-1}\right)$, while the red circle centered on the observation, Z_n, represents the observation model $p\left(z_n|a_n, \phi_n\right)$. To illustrate the figure, it can be seen that the product of the two

models gives

$$\begin{aligned} p\left(z_n, a_n, \phi_n | \mathcal{R}_{n-1}\right) &= p\left(a_n, \phi_n | \mathcal{R}_{n-1}\right) p\left(-w_n = a_n\phi_n - z_n | \mathcal{R}_{n-1}\right) \\ &= p\left(a_n, \phi_n | \mathcal{R}_{n-1}\right) p\left(z_n | a_n, \phi_n\right) \end{aligned} \tag{5.17}$$

where the second term is the distribution of $-W_n$ but offset by the observation Z_n, which is represented by the red circle in Figure 5.4.

Taking (5.16) into (5.12), a closed-form expression can be derived for the estimator (5.12) using [94, Eq. 6.643.2 and 9.220.2]

$$\tilde{\mu}_{n|n} = \int_0^\infty a_n p(a_n|\mathcal{R}_n) da_n \tag{5.18}$$

$$= \frac{\int_0^\infty \int_0^{2\pi} a_n^{2\gamma_n} \exp\left\{-\frac{a_n^2}{\beta_n^2} - \frac{1}{\nu_n^2}|z_n - a_n e^{j\phi_n}|^2\right\} d\phi_n da_n}{\int_0^\infty \int_0^{2\pi} a_n^{2\gamma_n - 1} \exp\left\{-\frac{a_n^2}{\beta_n^2} - \frac{1}{\nu_n^2}|z_n - a_n e^{j\phi_n}|^2\right\} d\phi_n da_n}$$

$$= \frac{\Gamma\left(\gamma_n + 0.5\right)}{\Gamma\left(\gamma_n\right)} \sqrt{\frac{\xi_n}{\zeta_n(\gamma_n + \xi_n)}} \frac{\mathcal{M}\left(\gamma_n + 0.5; 1; \frac{\zeta_n \xi_n}{\gamma_n + \xi_n}\right)}{\mathcal{M}\left(\gamma_n; 1; \frac{\zeta_n \xi_n}{\gamma_n + \xi_n}\right)} r_n \tag{5.19}$$

where r_n represents a realization of the random variable R_n, $\mathcal{M}$ is the confluent hypergeometric function [89], and

$$\xi_n = \frac{\mathrm{E}(A_n^2|\mathcal{R}_{n-1})}{\nu_n^2} = \frac{\tilde{\mu}_{n|n-1}^2 + \tilde{\sigma}_{n|n-1}^2}{\nu_n^2} = \frac{\gamma_n \beta_n^2}{\nu_n^2}, \ \zeta_n = \frac{r_n^2}{\nu_n^2} \tag{5.20}$$

are the a priori SNR and a posteriori SNR respectively. The details of the derivation of (5.19) is given in Section B.1 of Appendix B. The variance of the a posteriori

estimate is given by [94, Eq. 6.643.2 and 9.220.2]

$$\begin{aligned}\tilde{\sigma}^2_{n|n} &= \mathrm{E}\left(A_n^2|\mathcal{R}_n,\phi_n\right) - \left(\mathrm{E}\left(A_n|\mathcal{R}_n,\phi_n\right)\right)^2 \\ &= \frac{\int_0^\infty\int_0^{2\pi} a_n^{2\gamma_n+1}\exp\left\{-\frac{a_n^2}{\beta_n^2}-\frac{1}{\nu_n^2}|z_n-a_ne^{j\phi_n}|^2\right\}d\phi_n da_n}{\int_0^\infty\int_0^{2\pi} a_n^{2\gamma_n-1}\exp\left\{-\frac{a_n^2}{\beta_n^2}-\frac{1}{\nu_n^2}|z_n-a_ne^{j\phi_n}|^2\right\}d\phi_n da_n} - \left(\mu_{n|n}\right)^2 \qquad (5.21)\\ &= \frac{\gamma_n\xi_n}{\zeta_n(\gamma_n+\xi_n)}\frac{\mathcal{M}\left(\gamma_n+1;1;\frac{\zeta_n\xi_n}{\gamma_n+\xi_n}\right)}{\mathcal{M}\left(\gamma_n;1;\frac{\zeta_n\xi_n}{\gamma_n+\xi_n}\right)}R_n^2 - \left(\mu_{n|n}\right)^2 \qquad (5.22)\end{aligned}$$

which is derived in the same way as the derivation of (5.19).

5.2.5. Update of state vector

The final step is to update the entire state vector and the associated covariance matrix, $\tilde{\mathbf{s}}_{n|n}$ and $\widetilde{\mathbf{\Sigma}}_{n|n}$. Because the current element of the state vector has been estimated, if it can be decorrelated with the rest elements of the state vector, the whole state vector can then be updated based on the difference between the posterior and prior estimate. In order to decorrelate the current observation from the rest of the state vector, the covariance matrix $\widetilde{\mathbf{\Sigma}}_{n|n-1}$ is decomposed as

$$\widetilde{\mathbf{\Sigma}}_{n|n-1} = \begin{bmatrix} \sigma^2_{n|n-1} & \mathbf{g}_n^T \\ \mathbf{g}_n & \mathbf{G}_n \end{bmatrix}$$

where $\mathbf{g}_n$ is a $(p-1)$-dimensional vector and $\mathbf{G}_n$ is a $(p-1)\times(p-1)$ matrix. The state vector is now transformed as

$$\mathbf{t}_{n|n-1} = \mathbf{H}_n\tilde{\mathbf{s}}_{n|n-1} \qquad (5.23)$$

using the transformation matrix $\mathbf{H}_n = \begin{bmatrix} 1 & \mathbf{0}^T \\ -\frac{\mathbf{g}_n}{\sigma^2_{n|n-1}} & \mathbf{I} \end{bmatrix}$. The covariance matrix, $\mathbf{U}_{n|n-1}$, of the transformed state vector $\mathbf{t}_{n|n-1}$ is given by

$$\begin{aligned}\mathbf{U}_{n|n-1} = E\left(\mathbf{t}_{n|n-1}\mathbf{t}^T_{n|n-1}\right) &= \mathbf{H}_n\widetilde{\mathbf{\Sigma}}_{n|n-1}\mathbf{H}^T_n \\ &= \begin{bmatrix} \sigma^2_{n|n-1} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{G}_n - \sigma^{-2}_{n|n-1}\mathbf{g}_n\mathbf{g}^T_n \end{bmatrix}.\end{aligned}$$

It can be seen that the first element of $\mathbf{t}_{n|n-1}$ is equal to $\tilde{\mu}_{n|n-1}$ and uncorrelated with any of the other elements and is therefore distributed as $\mathcal{N}(\tilde{\mu}_{n|n-1}, \tilde{\sigma}^2_{n|n-1})$. Using the posterior mean and variance from (5.18) and (5.22), the transformed mean vector and covariance matrix can be updated as

$$\begin{aligned}\mathbf{z}_{n|n} &= \mathbf{z}_{n|n-1} + \left(\tilde{\mu}_{n|n} - \tilde{\mu}_{n|n-1}\right)\tilde{\mathbf{d}} \\ \mathbf{U}_{n|n} &= \mathbf{U}_{n|n-1} + \left(\tilde{\sigma}^2_{n|n} - \tilde{\sigma}^2_{n|n-1}\right)\tilde{\mathbf{d}}\tilde{\mathbf{d}}^T\end{aligned}$$

Inverting the transformation in (5.23), the following update equations can be obtained

$$\tilde{\mathbf{s}}_{n|n} = \tilde{\mathbf{s}}_{n|n-1} + \left(\tilde{\mu}_{n|n} - \tilde{\mu}_{n|n-1}\right)\left(\tilde{\sigma}^2_{n|n-1}\right)^{-1}\tilde{\mathbf{d}} \tag{5.24}$$

$$\widetilde{\mathbf{\Sigma}}_{n|n} = \widetilde{\mathbf{\Sigma}}_{n|n-1} + \left(\tilde{\sigma}^2_{n|n}\left(\tilde{\sigma}^2_{n|n-1}\right) - 1\right)\left(\tilde{\sigma}^2_{n|n-1}\right)^{-1}\widetilde{\mathbf{\Sigma}}_{n|n-1}\tilde{\mathbf{d}}\tilde{\mathbf{d}}^T\widetilde{\mathbf{\Sigma}}_{n|n-1} \tag{5.25}$$

In this section the update equations for the KF has been derived. For each acoustic frame of noisy speech, the a priori state vector $\tilde{\mathbf{s}}_{n|n-1}$ is first calculated and the corresponding covariance $\widetilde{\mathbf{\Sigma}}_{n|n-1}$, and solve (5.11) to find γ_n. (5.18) and (5.22) are then used to calculate the a posteriori estimate of the amplitude and the corresponding

variance respectively. Finally, the KF state vector and its covariance matrix are updated using (5.24) and (5.25).

5.2.6. Alternative Signal Addition Model

The enhancement algorithm described in Section 5.2.4 and (5.2.5) above differs from that proposed in [66] in two aspects: the use of generalized Gamma prior in (5.8) and the signal model in (5.1) that is additive in the complex STFT domain rather than the spectral amplitude domain. To asses the relative benefits of these two extensions, a version of our algorithm has also been implemented in which the generalized Gamma prior is used with a signal model that is additive in the spectral amplitude domain, i.e. $R_n = A_n + F_n$. Thus the model in (5.13) and now becomes

$$\begin{aligned} p\left(a_n|\mathcal{R}_n\right) &= p\left(a_n|r_n,\mathcal{R}_{n-1}\right) = \frac{p\left(a_n,r_n,\mathcal{R}_{n-1}\right)}{p\left(r_n,\mathcal{R}_{n-1}\right)} \\ &= \frac{p\left(r_n|a_n,\mathcal{R}_{n-1}\right)p\left(a_n|\mathcal{R}_{n-1}\right)p\left(\mathcal{R}_{n-1}\right)}{p\left(r_n|\mathcal{R}_{n-1}\right)p\left(\mathcal{R}_{n-1}\right)} \\ &= \frac{p\left(r_n|a_n,\mathcal{R}_{n-1}\right)p\left(a_n|\mathcal{R}_{n-1}\right)}{p\left(r_n|\mathcal{R}_{n-1}\right)}. \end{aligned} \tag{5.26}$$

Because R_n is conditionally independent of $\mathcal{R}_{n-1}$ given a_n, (5.26) becomes

$$p\left(a_n|\mathcal{R}_n\right) = \frac{p\left(r_n|a_n\right)p\left(a_n|\mathcal{R}_{n-1}\right)}{p\left(r_n|\mathcal{R}_{n-1}\right)}$$

Under the assumption that the signal model is additive in the spectral amplitude domain, the Gaussian observation prior model is

$$p\left(r_n|a_n\right) = \frac{1}{\sqrt{2\pi\nu_n^2}}\exp\left\{-\frac{\left(R_n-a_n\right)^2}{2\nu_n^2}\right\}$$

and the prior model of speech amplitude is also assumed to be the generalized

Gamma distribution in (5.8). Thus the posterior distribution of the speech amplitude is obtained as

$$\begin{aligned} p\left(a_n|\mathcal{R}_n\right) &= \frac{p\left(r_n|a_n\right)p\left(a_n|\mathcal{R}_{n-1}\right)}{p\left(r_n|\mathcal{R}_{n-1}\right)} \\ &= \frac{p\left(r_n|a_n\right)p\left(a_n|\mathcal{R}_{n-1}\right)}{\int_0^\infty p\left(r_n,a_n|\mathcal{R}_{n-1}\right)da_n} \\ &= \frac{p\left(r_n|a_n\right)p\left(a_n|\mathcal{R}_{n-1}\right)}{\int_0^\infty p\left(r_n|a_n\right)p\left(a_n|\mathcal{R}_{n-1}\right)da_n} \\ &= \frac{\frac{\sqrt{2}a_n^{2\gamma-1}}{\pi\nu_n^2\beta^{2\gamma}\Gamma(\gamma)}\exp\left\{-\frac{r_n^2}{2\nu_n^2}\right\}\exp\left\{-\left(\frac{1}{2\nu_n^2}+\frac{1}{\beta^2}\right)a_n^2+\frac{r_n}{\nu_n^2}a_n\right\}}{\int_0^\infty \frac{\sqrt{2}a_n^{2\gamma-1}}{\pi\nu_n^2\beta^{2\gamma}\Gamma(\gamma)}\exp\left\{-\frac{r_n^2}{2\nu_n^2}\right\}\exp\left\{-\left(\frac{1}{2\nu_n^2}+\frac{1}{\beta^2}\right)a_n^2+\frac{r_n}{\nu_n^2}a_n\right\}da_n} \end{aligned}$$

Thus the estimator of the amplitude (which is referred to as the "intermediate" estimator, KMMSEI) is given by

$$\begin{aligned} \tilde{\mu}_{n|n}^{(\mathrm{I})} &= \mathrm{E}\left(A_n|\mathcal{R}_n\right) = \int_0^\infty a_n p(a_n|\mathcal{R}_n)da_n \\ &= \frac{\int_0^\infty \frac{\sqrt{2}a_n^{2\gamma_n}}{\pi\nu_n^2\beta^{2\gamma}\Gamma(\gamma_n)}\exp\left\{-\frac{r_n^2}{2\nu_n^2}\right\}\exp\left\{-\left(\frac{1}{2\nu_n^2}+\frac{1}{\beta^2}\right)a_n^2+\frac{r_n}{\nu_n^2}a_n\right\}da_n}{\int_0^\infty \frac{\sqrt{2}a_n^{2\gamma_n-1}}{\pi\nu_n^2\beta^{2\gamma}\Gamma(\gamma_n)}\exp\left\{-\frac{r_n^2}{2\nu_n^2}\right\}\exp\left\{-\left(\frac{1}{2\nu_n^2}+\frac{1}{\beta^2}\right)a_n^2+\frac{r_n}{\nu_n^2}a_n\right\}da_n} \end{aligned} \tag{5.27}$$

and the closed-form of the intermediate estimator can be obtained using [94, Eq. 3.462.1] as

$$\tilde{\mu}_{n|n}^{(\mathrm{I})} = -\left(\frac{1}{2\nu_n^2}+\frac{1}{\beta_n^2}\right)^{-\frac{1}{2}}\frac{\Gamma\left(2\gamma_n+1\right)}{\Gamma\left(2\gamma_n\right)}\frac{\mathcal{D}_{-2\gamma_n-1}\left(-\frac{r_n}{\nu_n^2\sqrt{2\left(\frac{1}{2\nu_n^2}+\frac{1}{\beta^2}\right)}}\right)}{\mathcal{D}_{-2\gamma_n}\left(-\frac{r_n}{\nu_n^2\sqrt{2\left(\frac{1}{2\nu_n^2}+\frac{1}{\beta^2}\right)}}\right)} \tag{5.28}$$

where $\mathcal{D}\left(\cdot\right)$is the parabolic cylinder function, the definition of which is given in Section A.2 of Appendix A.

By substituting the calculation of the a priori SNR and a posteriori SNR in (5.20) into (5.28), it becomes

$$\tilde{\mu}_{n|n}^{(\mathrm{I})} = \frac{2\gamma_n}{\zeta_n}\sqrt{\frac{\xi_n\zeta_n}{\xi_n+2\zeta_n}}\frac{\mathcal{D}_{-2\gamma_n-1}\left(-\sqrt{\frac{\xi_n\zeta_n}{\xi_n+2\zeta_n}}\right)}{\mathcal{D}_{-2\gamma_n-1}\left(-\sqrt{\frac{\xi_n\zeta_n}{\xi_n+2\zeta_n}}\right)}r_n \tag{5.29}$$

The corresponding variance of the estimate is given by [94, Eq. 3.462.1]

$$\begin{aligned}
\tilde{\sigma}_{n|n}^{2(\mathrm{I})} &= \mathrm{E}\left(A_n^2|\mathcal{R}_n\right) - \left(\mathrm{E}\left(A_n|\mathcal{R}_n\right)\right)^2 \\
&= \frac{\int_0^\infty a_n^2 p\left(r_n|a_n\right) p\left(a_n|\mathcal{R}_{n-1}\right) da_n}{\int_0^\infty p\left(r_n, a_n|\mathcal{R}_{n-1}\right) da_n} - \left(\tilde{\mu}_{n|n}^{\mathrm{I}}\right)^2 \\
&= \frac{\int_0^\infty \frac{\sqrt{2}a_n^{2\gamma_n+1}}{\pi\nu_n^2\beta^{2\gamma_n}\Gamma(\gamma_n)}\exp\left\{-\frac{r_n^2}{2\nu_n^2}\right\}\exp\left\{-\left(\frac{1}{2\nu_n^2}+\frac{1}{\beta^2}\right)a_n^2+\frac{r_n}{\nu_n^2}a_n\right\}da_n}{\int_0^\infty \frac{\sqrt{2}a_n^{2\gamma_n-1}}{\pi\nu_n^2\beta^{2\gamma_n}\Gamma(\gamma_n)}\exp\left\{-\frac{r_n^2}{2\nu_n^2}\right\}\exp\left\{-\left(\frac{1}{2\nu_n^2}+\frac{1}{\beta^2}\right)a_n^2+\frac{r_n}{\nu_n^2}a_n\right\}da_n} - \left(\tilde{\mu}_{n|n}^{\mathrm{I}}\right)^2 \\
&= \frac{2\left(2\gamma_n+1\right)}{\zeta_n^2}\frac{\xi_n\zeta_n}{\xi_n+2\zeta_n}\frac{\mathcal{D}_{-2\gamma_n-2}\left(-\sqrt{\frac{\xi_n\zeta_n}{\xi_n+2\zeta_n}}\right)}{\mathcal{D}_{-2\gamma_n}\left(-\sqrt{\frac{\xi_n\zeta_n}{\xi_n+2\zeta_n}}\right)}r_n^2 - \left(\tilde{\mu}_{n|n}^{\mathrm{I}}\right)^2
\end{aligned} \tag{5.30}$$

which is derived in the same way as the derivation of (5.29).

5.2.7. Implementation and evaluation

In this section, the performance of six enhancement algorithms is compared:

(i) logMMSE – the baseline enhancer from [53, 85] that is introduced in Section 2.4;

(ii) Perceptual Motivated MMSE (pMMSE) – the MMSE estimator from [58, 85] that is introduced in Section 2.4 using a weighted Euclidean distortion measure with a power exponent of $u = -1$ in (2.14);

(iii) MDST – the enhancer from [65] that is introduced in Section 2.5.2;

(iv) Modulation Domain Kalman filter that assumes white noise (MDKF) – the version of the modulation-domain Kalman filter from [66] that assumes white noise

and that extracts the modulation-domain LPC coefficients from enhanced speech (using the logMMSE algorithm [53, 85]);

(v) Kalman filter based MMSE estimator (KMMSE) – the proposed enhancer described in Section 5.2.4 and 5.2.5 that uses a generalized Gamma prior for speech spectral amplitudes and a signal model that is additive in the complex STFT domain.

(vi) Intermediate KMMSE (KMMSEI) – the intermediate version of our proposed algorithm that combines a generalized Gamma prior for the speech spectral amplitudes with a signal model that is additive in the spectral amplitude domain. The details of derivation of this estimator are described in Section 5.2.6.

The parameters of all the algorithms were chosen to optimize performance on the development set described in Section 1.4.1.1. The sensitivity of the orders of LPC models of speech and noise has been discussed in Section 1.4.1.4.

Parameter	Settings
Sampling frequency	8 kHz
Acoustic frame length	16 ms
Acoustic frame increment	4 ms
Modulation frame length	64 ms
Modulation frame increment	16 ms
Analysis-synthesis window	Hamming window
Speech LPC model order p	3

Table 5.1.: Parameter setting in experiments.

In the experiments, the core test set from the TIMIT database (for details see Chapter 2) is used and the speech is corrupted by the noise from the RSG-10 database [25] and the ITU-T test signals database [87] at $-10, -5, 0, 5, 10$ and 15 dB global SNR. The noise power spectrum, $\nu^2_{n,k}$, is estimated using the algorithm from [43] as implemented in [85] and it is used in logMMSE, pMMSE, MDKF, KMMSE and KMMSEI algorithms.

The performance of the algorithms is evaluated using both segmental SNR (segSNR) and the Perceptual Evaluation of Speech Quality (PESQ) measure. All the measured values shown are averages over all the sentences in the TIMIT core test set. Figure 5.5 and 5.6 show respectively the average segSNR of speech enhanced by the proposed algorithm (KMMSE) as well as by the logMMSE, pMMSE and MDKF algorithms for car noise [25] and street noise [87]. It shows that for car noise, which is predominantly low frequency, pMMSE gives the best segSNR especially at poor SNRs where it is approximately 2 dB better than KMMSE, the next best algorithm. For street noise however, which has a broader spectrum, the situation is reversed and the KMMSE algorithm has the best performance especially at SNRs above 5 dB. Figure 5.7 and 5.8 show the corresponding average PESQ scores for car noise and street noise, respectively. It can be seen that, with this measure, the KMMSE algorithm clearly has the highest performance. For car noise, the PESQ score from the KMMSE algorithm is approximately 0.2 better than that of the other algorithms at SNRs below 5 dB while for street noise, the corresponding figure is 0.15. These differences correspond to SNR improvements of 4 dB and 2.5 dB respectively. To assess the robustness to noise type, the algorithms has been evaluated using twelve different noise types from [25] with the average SNR for each noise type chosen to give a mean PESQ score of 2.0 for the noisy speech. In 5.9, the solid lines show the median, the boxes the interquartile range and the whiskers the extreme PESQ values for the 198×12 speech-plus-noise combinations. Figure 5.10 shows box plots of the difference in PESQ score between competing algorithms and KMMSE. It shows that in all cases the entire box lies below the axis line; this indicates that KMMSE results in an improvement for an overwhelming majority of speech-plus-noise combinations. The KMMSEI box plot demonstrates the small but consistent benefit of using an additive model in the complex STFT domain rather than the

amplitude domain.

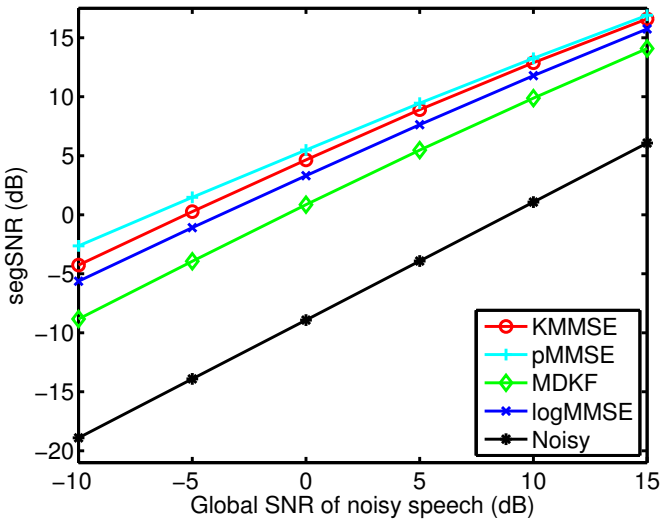

Figure 5.5.: Average segmental SNR of enhanced speech after processing by four algorithms plotted against the global SNR of the input speech corrupted by additive car noise.

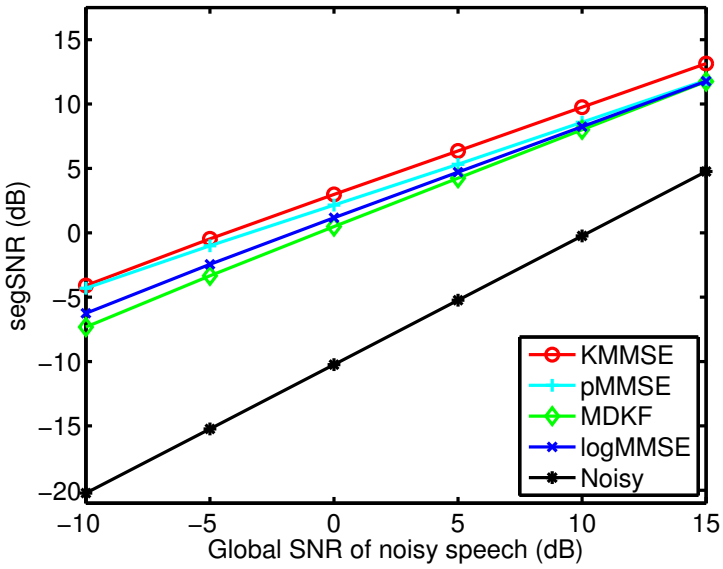

Figure 5.6.: Average segmental SNR of enhanced speech after processing by four algorithms plotted against the global SNR of the input speech corrupted by additive street noise

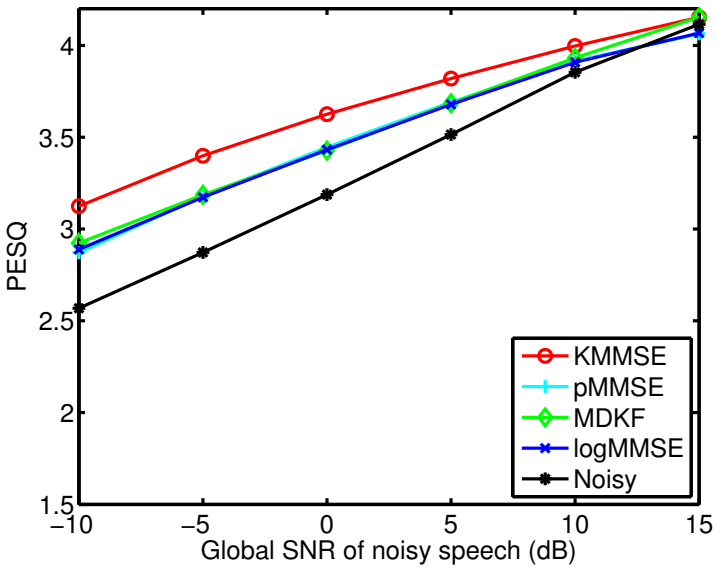

Figure 5.7.: Average PESQ quality of enhanced speech after processing by four algorithms plotted against the global SNR of the input speech corrupted by additive car noise

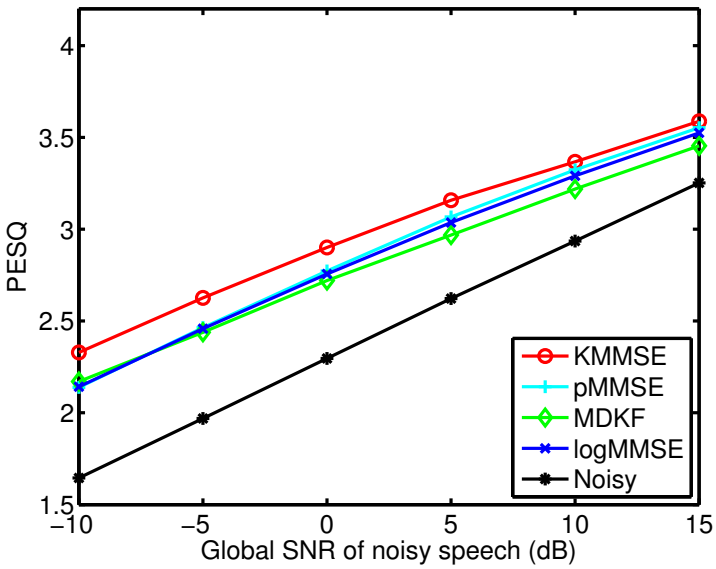

Figure 5.8.: Average PESQ quality of enhanced speech after processing by four algorithms plotted against the global SNR of the input speech corrupted by additive street noise

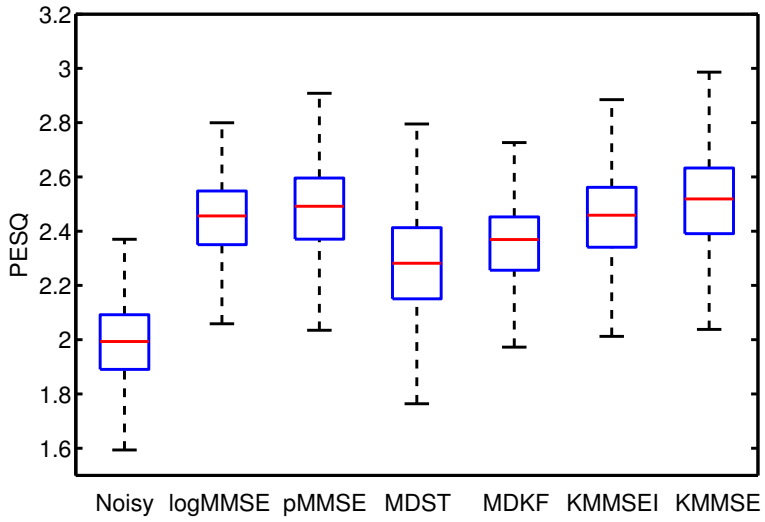

Figure 5.9.: Box plot of the PESQ scores for noisy speech processed by six enhancement algorithms. The plots show the median, interquartile range and extreme values from 2376 speech+noise combinations.

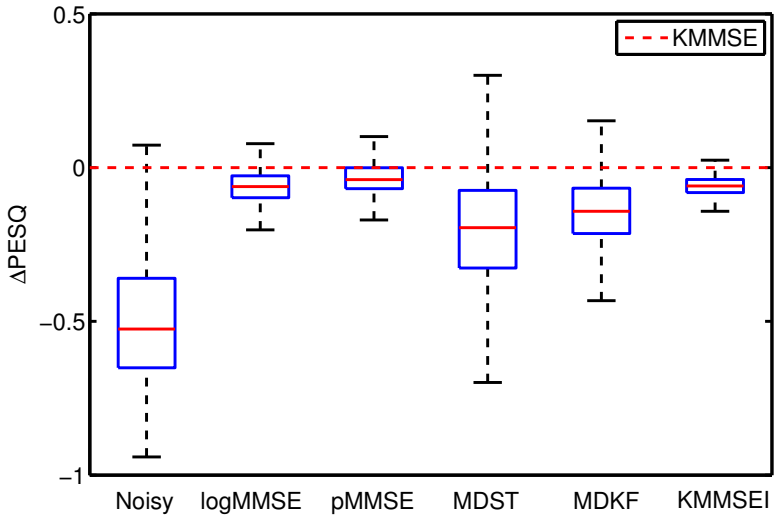

Figure 5.10.: Box plot showing the difference in PESQ score between competing algorithms and the proposed algorithm, KMMSE for 2376 speech+noise combinations.

5.3. Enhancement with Gaussring priors

In deriving the KMMSE enhancement algorithm in Section 5.2, the noise was assumed to be stationary and the Kalman filter tracked only the speech dynamics. However, within the Kalman filter, it is possible to include the noise dynamics as well, as was done in Chapter 3. The state vector of speech, $\tilde{\mathbf{s}}$, and noise, $\check{\mathbf{s}}$, are concatenated to form a single state vector $\mathbf{s}$, and in the Kalman filtering the entire state vector is estimated and propagated. The equations for estimating $\mathbf{s}$ have been given in (2.20) to (2.24). For this case, the observation model, $|Z_{n,k}| = |S_{n,k}| + |W_{n,k}|$, can be seen as a constraint applied to the speech and noise when deriving the MMSE estimate for their amplitudes. In this section, as in Section 5.2, the speech and noise are also assumed to be additive in the complex STFT domain and the speech and noise STFT coefficients are assumed to have uniform prior phase distributions. To derive the Kalman filter update, the mean and variance need to estimate. However, in this case the denominator in (5.13) is now calculated as

$$\begin{aligned} &p\left(z_n|\mathcal{R}_{n-1}, \mathcal{F}_{n-1}\right) \\ &= \int_0^\infty \int_0^{2\pi} \int_0^\infty \int_0^{2\pi} p\left(z_n|a_n, \phi_n, f_n, \psi_n\right) p\left(f_n, \psi_n, a_n, \phi_n, |\mathcal{R}_{n-1}, \mathcal{F}_{n-1}\right) d\phi_n da_n d\psi_n df_n \end{aligned} \tag{5.31}$$

where $\mathcal{F}_n = [F_1 \dots F_n]$ represents the observed noise amplitudes up to time n. The derivation of the MMSE estimator is also under the constraint in (5.1). This derivation is mathematically intractable if a generalized Gamma distribution, as used in Section 5.2, is assumed for both the speech and noise prior amplitude distributions.

In order to overcome the above problem, in this section a distribution, "Gaussring", is proposed for the complex STFT coefficients that comprises a mixture of Gaussians whose centres lie in a circle on the complex plane.

5.3.1. Gaussring properties

5.3.1.1. Gaussring distribution

From the colored noise version modulation-domain Kalman filter described in Section 2.5.1, the prior estimate of the amplitude of both speech and noise can be obtained. The real and imaginary part of complex STFT coefficients of clean speech are denoted as $\tilde{r}_{n|n-1}$ and $\tilde{i}_{n|n-1}$ respectively, and those of noise as $\breve{r}_{n|n-1}$ and $\breve{i}_{n|n-1}$ respectively. The idea of the Gaussring model is, under the assumption that the phase of the complex STFT coefficients of speech and noise is uniformly distributed, to use a mixture of 2-dimensional circular Gaussians with the uniform weight to approximate the joint prior distribution $p(\tilde{r}_{n|n-1}, \tilde{i}_{n|n-1})$ and $p(\breve{r}_{n|n-1}, \breve{i}_{n|n-1})$. Without loss of generality, in this and the following subsections the distribution of speech or noise will be denoted as $p\left(r_{n|n-1}, i_{n|n-1}\right)$.

The Gaussring model is defined as

$$p(r_{n|n-1}, i_{n|n-1}) = \sum_{j=1}^{J} \epsilon_{n|n-1}^{(j)} \mathcal{N}\left(\boldsymbol{\mu}_{n|n-1}^{(j)}, \boldsymbol{\Sigma}_{n|n-1}^{(j)}\right), \tag{5.32}$$

where J is the number of the mixtures for the speech and noise, respectively. The 2-dimensional mean vector $\boldsymbol{\mu}_{n|n-1}^{(j)}$and the 2×2 covariance matrix $\boldsymbol{\Sigma}_{n|n-1}^{(j)}$ of each mixture are given by

$$\boldsymbol{\mu}_{n|n-1}^{(j)} = \begin{bmatrix} \mu_r^{(j)} & \mu_i^{(j)} \end{bmatrix}^T \tag{5.33}$$

$$\boldsymbol{\Sigma}_{n|n-1}^{(j)} = \begin{bmatrix} \sigma_r^{2(j)} & 0 \\ 0 & \sigma_i^{2(j)} \end{bmatrix}. \tag{5.34}$$

Because each Gaussian is circular on the complex plane, the variance of the real part, $\sigma_r^{2(j)}$, and that of the imaginary part, $\sigma_i^{2(j)}$, are equal. Therefore both of them

can be denoted as $\sigma^{2(j)}$. In order to fit the ring distribution obtained from the prior estimate, the number of the Gaussian components (circles), J, depends on the ratio of the mean and standard deviation of the prior estimate and is chosen so that the mixture centres are separated by 2σ around a circle of radius μ; this gives

$$J = \left\lceil \frac{\pi \mu_{n|n-1}}{\sigma_{n|n-1}} \right\rceil$$

where $\lceil \cdot \rceil$ is the ceiling function. When $\sigma_{n|n-1}$ is much larger than $\mu_{n|n-1}$, a minimum value 3 for J is set to ensure that the phase is uniformly distributed. Thus J is set to be

$$J = \max\left(\left\lceil \frac{\pi \mu_{n|n-1}}{\sigma_{n|n-1}} \right\rceil, 3 \right) \tag{5.35}$$

The examples of Gaussring models matching the prior estimate are given from Figure 5.11 to Figure 5.13 for $\mu_{n|n-1} = 2,\ 10,\ 1, 0.1$ and $\sigma_{n|n-1} = 1$. In these cases, the models are assumed to be centered at the origin. The marginal amplitude distribution (Rician) and phase distribution (uniform) of the Gaussring model are also shown on the right of the figures. The white circles shown in the complex plane represent the mean of each Gaussian component. For Rician distribution, the mean μ_{Rician} and standard deviation σ_{Rician} satisfies

$$\frac{\sigma_{Rician}}{\mu_{Rician}} \leq \sqrt{\frac{4}{\pi} - 1} \tag{5.36}$$

and when $\frac{\sigma_{Rician}}{\mu_{Rician}} = \sqrt{\frac{4}{\pi} - 1}$, it becomes Rayleigh distribution. As a result, when $\frac{\sigma_{Rician}}{\mu_{Rician}} > \sqrt{\frac{4}{\pi} - 1}$, the actual fitted mean and standard deviation deviate from the actual values, which can be seen in Figures 5.13 and 5.14. In these cases, the model will be fitted with a mean and standard deviation which obey the inequality. In the

Kalman filtering, the constraint in (5.36) is placed on the prior estimate.

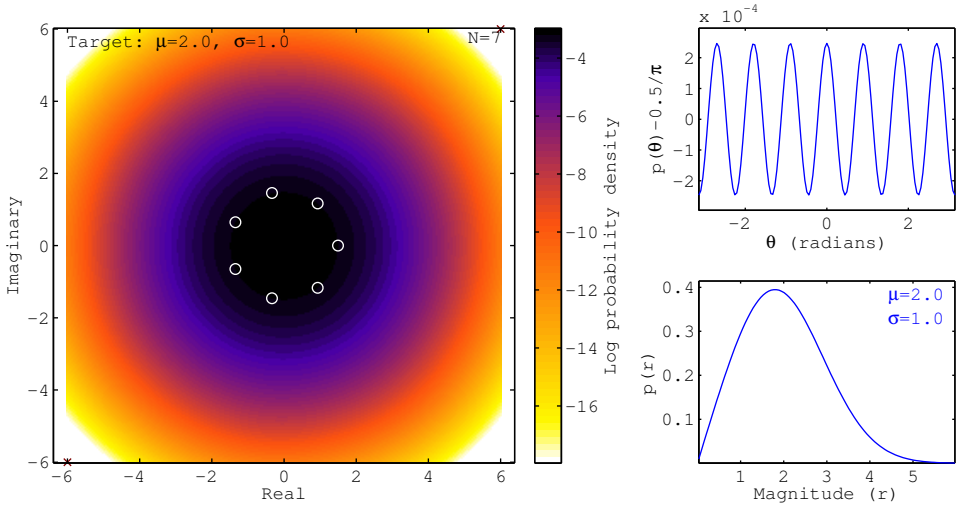

Figure 5.11.: Gaussring model fit for $\mu_{n|n-1} = 2$ and $\sigma_{n|n-1} = 1$.

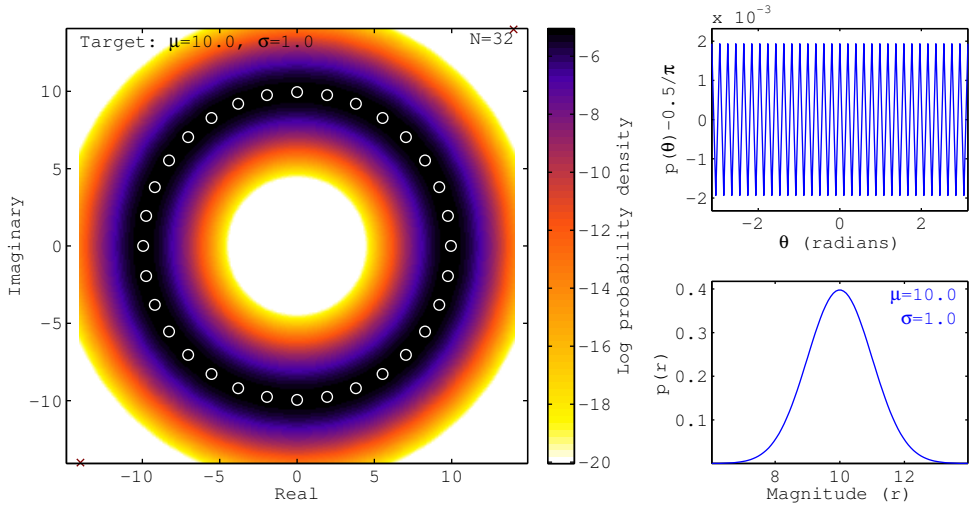

Figure 5.12.: Gaussring model fit for $\mu_{n|n-1} = 10$ and $\sigma_{n|n-1} = 1$.

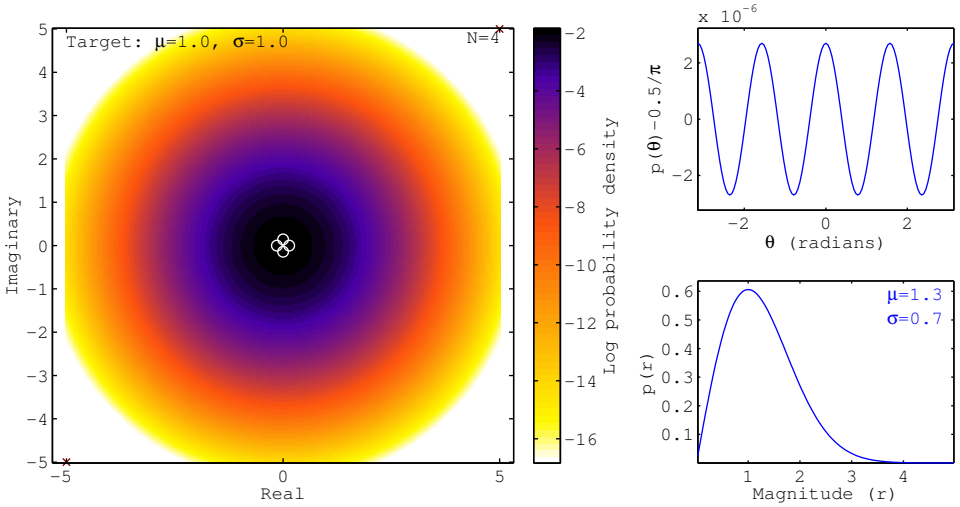

Figure 5.13.: Gaussring model fit for $\mu_{n|n-1} = 1$ and $\sigma_{n|n-1} = 1$.

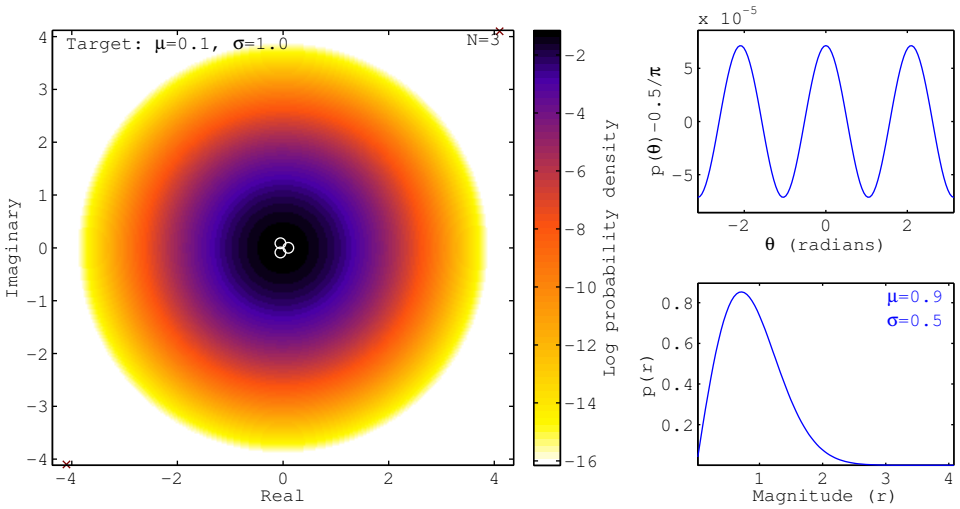

Figure 5.14.: Gaussring model fit for $\mu_{n|n-1} = 0.9$ and $\sigma_{n|n-1} = 0.5$.

5.3.1.2. Posterior distribution

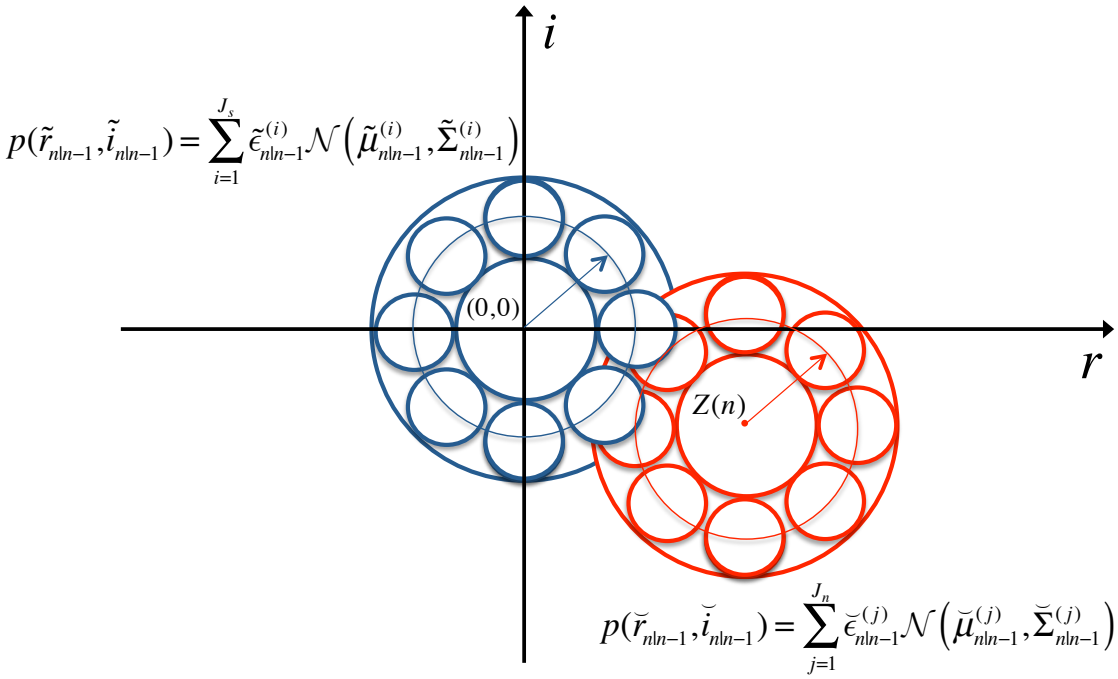

Figure 5.15.: Gaussring model of speech and noise. Blue circles represent the speech Guassring model and red circles represent the noise Guassring model.

Using the Gaussring distribution, a mixture of Gaussians can be fit for both the speech and noise prior estimates. An example showing the combination of the Gaussring models of the speech and noise is given in Figure 5.15. To guarantee that the sum of the speech and noise in the complex STFT domain is the STFT coefficients of the noisy speech, the Gaussring of speech is assumed to be centered at the original and that of the noise is centered at the observation Z_n. As shown in (5.17), the posterior distribution is calculated as a product of the each pair of the Gaussian components of speech and noise, which is normalized by a factor to make the sum of the posterior distribution equal to 1. Thus, supposing there are J_s Gaussian components for the speech and J_n Gaussian components for the noise, a total of $J_s J_n$ Gaussian components will be obtained for the posterior distribution

after combining the speech and noise prior model. The product of the ith component of speech and jth component of noise, denoted as $\mathcal{N}\left(\boldsymbol{\mu}_{n|n}^{(ij)}, \boldsymbol{\Sigma}_{n|n}^{(ij)}, \epsilon_{n|n}^{(ij)}\right)$, is calculated as [83]

$$\boldsymbol{\Sigma}_{n|n}^{(ij)} = \left(\left(\widetilde{\boldsymbol{\Sigma}}_{n|n-1}^{(i)}\right)^{-1} + \left(\breve{\boldsymbol{\Sigma}}_{n|n-1}^{(j)}\right)^{-1}\right)^{-1} \tag{5.37}$$

$$\boldsymbol{\mu}_{n|n}^{(ij)} = \boldsymbol{\Sigma}_{n|n}^{(i)}\left(\tilde{\boldsymbol{\mu}}_{n|n-1}^{(i)}\left(\widetilde{\boldsymbol{\Sigma}}_{n|n-1}^{(i)}\right)^{-1} + \breve{\boldsymbol{\mu}}_{n|n-1}^{(j)}\left(\breve{\boldsymbol{\Sigma}}_{n|n-1}^{(j)}\right)^{-1}\right) \tag{5.38}$$

$$\epsilon_{n|n}^{(ij)} = \tilde{\epsilon}_{n|n-1}^{(i)}\mathcal{N}\left(\tilde{\boldsymbol{\mu}}_{n|n-1}^{(i)}, \widetilde{\boldsymbol{\Sigma}}_{n|n-1}^{(i)}\right)\breve{\epsilon}_{n|n-1}^{(j)}\mathcal{N}\left(\breve{\boldsymbol{\mu}}_{n|n-1}^{(j)}, \breve{\boldsymbol{\Sigma}}_{n|n-1}^{(j)}\right) \tag{5.39}$$

The optimal estimate of the amplitude of speech and noise is calculated as the mean of the amplitude of posterior Gaussians as in (5.12). In the next subsections, how the parameters of the Gaussring model are estimated by matching the moment of the prior estimate will be described. Also, the calculation of the optimal estimate of the amplitude, its variance and the covariance of the amplitudes of speech and noise will be introduced.

5.3.2. Moment Matching

5.3.2.1. Amplitude distribution

Because each mixture component in the Gaussring model is a circular Gaussian, its amplitude is Rician distributed [89] which is a 2-parameter distribution given by

$$p\left(x; \upsilon, \alpha\right) = \frac{x}{\upsilon^2}\exp\left(\frac{-(x^2+\alpha^2)}{2\upsilon^2}\right) I_0\left(\frac{x\alpha}{\upsilon^2}\right)$$

where $I_0\left(\cdot\right)$ is the modified Bessel function of the first kind with order 0. Thus, the parameters of the Rician distribution given the prior estimate of the modulation-domain Kalman filter need to be estimated.. The mean and variance of the Rician

distribution is given by

$$\mu_{Rician} = \upsilon\sqrt{\frac{\pi}{2}}\exp\left(-\frac{\alpha^2}{2\upsilon^2}\right)\left[\left(1-\frac{\alpha^2}{2\upsilon^2}\right)I_0\left(-\frac{\alpha^2}{4\upsilon^2}\right)-\frac{\alpha^2}{2\upsilon^2}I_1\left(-\frac{\alpha^2}{4\upsilon^2}\right)\right] \tag{5.40}$$
$$\sigma^2_{Rician} = 2\upsilon^2+\alpha^2-\mu^2_{Rician}$$

where $\alpha \geq 0$ and $\upsilon \geq 0$ are the parameters of the Rician distribution. I_1 is the modified Bessel function of the first kind with 1. It can be seen that it is difficult to invert (5.40) to determine α and υ from the prior mean $\mu_{n|n-1}$ and variance $\sigma^2_{n|n-1}$ because Bessel functions are involved in the calculation of the mean and variance. In this section an efficient method to estimate the parameters of the Rician distribution from the prior estimate will be introduced.

The Nakagami-m distribution [95] is used to approximate the Rician distribution. The mean and variance of the Nakagami-m distribution are given by

$$\mu_{nakagami} = \frac{\Gamma(m+\frac{1}{2})}{\Gamma(m)}\left(\frac{\Omega}{m}\right)^{1/2} \tag{5.41}$$
$$\sigma^2_{nakagami} = \Omega\left(1-\frac{1}{m}\left(\frac{\Gamma(m+\frac{1}{2})}{\Gamma(m)}\right)^2\right) \tag{5.42}$$

where $\Omega > 0$ and $m > 0.5$ are the parameters of the distribution. The Nakagami-m distribution is an accurate approximation of the Rician distribution when the parameter in the Nakagami-m distribution $m > 1$ [96, 97, 98]. When $0.5 < m < 1$, Nakagami-m distribution is an accurate approximation of the Hoyt distribution [99], which is a distribution for modeling the amplitude of complex-valued signals whose real and imaginary parts have zero mean and unequal variance. Because the real and imaginary parts of the complex STFT coefficients of both speech and noise have equal variance, the Nakagami-m distribution can be used to approximate the Rician distribution. The parameters of the Rician distribution can be obtained from

the parameters of the corresponding Nakagami-m distribution by moment matching [98], which is given by

$$\frac{2\upsilon^2}{\alpha^2} = \frac{m - \sqrt{m^2 - m}}{\sqrt{m^2 - m}} \triangleq \Lambda\,(m > 1) \tag{5.43}$$

$$2\upsilon^2 + \alpha^2 = \Omega \tag{5.44}$$

The equations for calculating the Rician distribution parameters from the corresponding Nakagami-m distributions can be obtained:

$$\alpha = \sqrt{\Omega/(1+\Lambda)} \tag{5.45}$$

$$\upsilon^2 = \alpha^2\Lambda/2 \tag{5.46}$$

In Figure 5.16, the Rician distribution and Nakamai-m distribution are compared for $\Omega = 0.1, 1, 10$ and $m = 2$, and the parameters of Rician distribution are calculated using (5.45) and (5.46). As shown in the graph, the Nakagami-m distribution is a close approximation of the Rician distribution for a range of parameters.

Thus, the Nakagami-m distribution can be used to model the amplitudes. As will be seen, there are two advantages of using the Nakagami distribution: Firstly the parameters of the distribution can be estimated efficiently by moment matching of the prior estimate; Secondly, the covariance of the amplitudes of the speech and noise can be approximated efficiently.

From (5.41) and (5.42), it can be seen that it is not directly tractable to obtain m and Ω from the prior estimate. However, because [97]

$$\sqrt{m - 1/4} < \frac{\Gamma(m + \frac{1}{2})}{\Gamma(m)} < m/\sqrt{m + 1/4}, \tag{5.47}$$

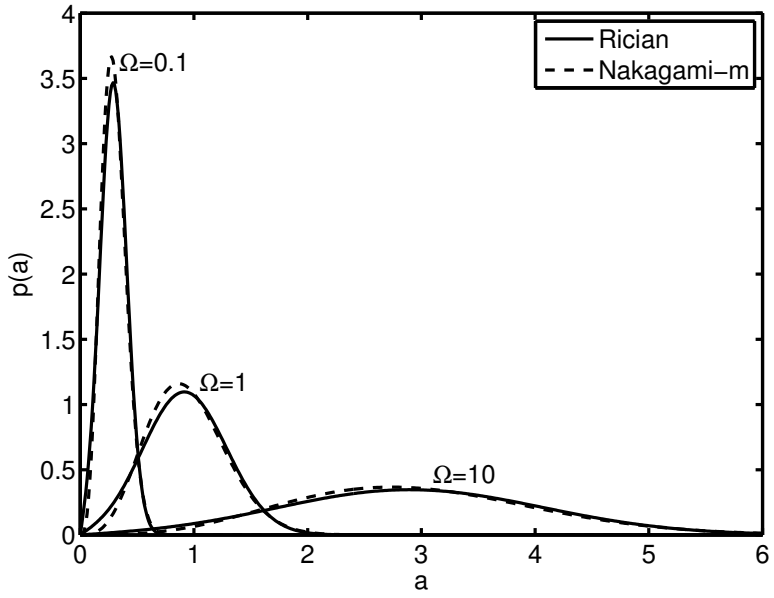

Figure 5.16.: Comparison of Rician and Nakagami distribution for $\Omega = 0.1, 1, 10$ and $m = 2$.

and the difference between the upper and lower boundaries satisfies

$$m/\sqrt{m+1/4} - \sqrt{m-1/4} < 0.077$$

The lower bounds in (5.47) can be used to approximate the ratio $\frac{\Gamma(m+\frac{1}{2})}{\Gamma(m)}$ and it has been shown in [97] that, when using the low bound $\sqrt{m-1/4}$, the resultant estimators based on the Nakagami-m posterior model are close approximations of these based on the Rician posterior model. Using the lower bound for the ratio $\frac{\Gamma(m+\frac{1}{2})}{\Gamma(m)}$ in (5.47), the parameters m and Ω can be calculated as

$$\Omega = \mu_{n|n-1}^2 + \sigma_{n|n-1}^2 \tag{5.48}$$

$$m = \Omega/(4\sigma_{n|n-1}^2). \tag{5.49}$$

Thus, the parameters of the corresponding Rician distribution from can be calculated

from Ω and m using (5.45) and (5.46). From the parameters α and υ^2, the mean and covariance of each components of the Gaussring model in (5.33) and (5.34) can be obtained as

$$\mu_r^{(j)} = \alpha \cos\left(2\pi j/J\right) \tag{5.50}$$

$$\mu_i^{(j)} = \alpha \sin\left(2\pi j/J\right) \tag{5.51}$$

$$\sigma^{2(j)} = \upsilon^2 \tag{5.52}$$

In the next subsection, the posterior estimate of the amplitudes and speech and noise, and also their covariance will be described.

5.3.2.2. Mean and covariance of amplitude

For each mixture of the posterior distribution, $\mathcal{N}(\boldsymbol{\mu}_{n|n}^{(j)}, \boldsymbol{\Sigma}_{n|n}^{(j)})$, the parameters of its amplitude, Ω and m, need to calculate. The real and imaginary part of the complex STFT coefficients of speech and noise are denoted as R and I, respectively. The first step is to calculate the mean and variance of the squared amplitude, which are defined here as $\mu_{ss}^{(j)} \triangleq \mathrm{E}\left(R_{n|n}^{2(j)} + I_{n|n}^{2(j)}\right)$ and $\sigma_{ss}^{2(j)} \triangleq \mathrm{Var}\left(R_{n|n}^{2(j)} + I_{n|n}^{2(j)}\right)$ respectively. The notation ss indicates sum of the squared real and imaginary parts. Supposing that the origin of the Gaussring model in the complex domain is (r_0, i_0), for each

mixture

$$\begin{aligned}\mu_{ss}^{(j)} &= \mathrm{E}\left(R_{n|n}^{2(j)}\right) + \mathrm{E}\left(I_{n|n}^{2(j)}\right) \\ &= \mu_r^{2(j)} + \sigma_r^{2(j)} + \mu_i^{2(j)} + \sigma_i^{2(j)} - 2\left(r_0\mu_r^{(j)} + i_0\mu_i^{(j)}\right) + \left(r_0^2 + i_0^2\right) \qquad (5.53)\\ \sigma_{ss}^{2(j)} &= \mathrm{E}\left(\left(R_{n|n}^{2(j)} + I_{n|n}^{2(j)}\right)^2\right) - \mu_{ss}^{2(j)} \\ &= \sigma_r^{2(j)}\left(2\sigma_r^{2(j)} + 4\mu_r^{2(j)}\right) + \sigma_i^{2(j)}\left(2\sigma_i^{2(j)} + 4\mu_i^{2(j)}\right) - ... \\ &\quad 8(r_0\sigma_r^{2(j)}\mu_r^{(j)} + i_0\sigma_i^{2(j)}\mu_i^{(j)}) + 4\left(r_0^2\sigma_r^{2(j)} + i_0^2\sigma_i^{2(j)}\right) \qquad (5.54)\end{aligned}$$

The second step is to obtain the parameters of the amplitude distribution of each mixture, $p_j(a_n)$,

$$\Omega^{(j)} = \mu_{ss}^{2(j)} \qquad (5.55)$$

$$m^{(j)} = \Omega^{2(j)}/\sigma_{ss}^{2(j)} \qquad (5.56)$$

Thus, the mean of the amplitude can be calculated as

$$\mu_a^{(j)} \triangleq \mathrm{E}\left(\sqrt{R_{n|n}^{2(j)} + I_{n|n}^{2(j)}}\right) = \frac{\Gamma\left(m^{(j)} + 0.5\right)}{\Gamma\left(m^{(j)}\right)}\sqrt{\frac{\Omega^{(j)}}{m^{(j)}}} \qquad (5.57)$$

After the mean, $\mu_a^{(j)}$, is obtained, the third step is to calculate the overall mean and variance of the a posteriori estimate of the amplitude. Because every mixture is equally weighted, the overall mean and variance are given by

$$\mu_{n|n} \triangleq \mathrm{E}\left(\sqrt{R_{n|n}^{2(j)} + I_{n|n}^{2(j)}}\right) = \sum_{j=1}^{J_sJ_n} \epsilon_{n|n}^{(j)}\mu_a^{(j)} \qquad (5.58)$$

$$\sigma_{n|n}^2 \triangleq \mathrm{Var}\left(\sqrt{R_{n|n}^{2(j)} + I_{n|n}^{2(j)}}\right) = \sum_{j=1}^{J_sJ_n} \epsilon_{n|n}^{(j)}\Omega^{(j)} - \mu_{n|n}^2 \qquad (5.59)$$

Defining the posterior estimate of the amplitude of the speech and noise as, $\widetilde{A}_{n|n} \triangleq \sqrt{\widetilde{R}^{2(j)}_{n|n} + \widetilde{I}^{2(j)}_{n|n}}$and $\breve{A}_{n|n} \triangleq \sqrt{\breve{R}^{2(j)}_{n|n} + \breve{I}^{2(j)}_{n|n}}$, respectively. The problem remains unsolved so far is the calculation of the covariance for the speech and noise amplitude, $\mathrm{E}\left(\widetilde{A}_{n|n}\breve{A}_{n|n}\right) - \mathrm{E}\left(\widetilde{A}_{n|n}\right)\mathrm{E}\left(\breve{A}_{n|n}\right)$. For two Nakagami-m variables with different m, there is no analytical solution for calculating the correlation coefficient, ρ_a, between the variables. ρ_a is defined as

$$\rho_a = \frac{\mathrm{E}\left(\widetilde{A}_{n|n}\breve{A}_{n|n}\right) - \mathrm{E}\left(\widetilde{A}_{n|n}\right)\mathrm{E}\left(\breve{A}_{n|n}\right)}{\sqrt{\mathrm{Var}\left(\widetilde{A}_{n|n}\right)\mathrm{Var}\left(\breve{A}_{n|n}\right)}} \tag{5.60}$$

To calculate ρ_a using (5.60) is not trivial. However, it is found that ρ_a can be approximated accurately by the correlation coefficient between the squared Nakagami-m variables [100], which is denoted as ρ_γ. Based on this observation, ρ_γ can be calculated instead of ρ_a. Under the assumption that the speech distribution is centered on the origin (0, 0) and the noise distribution is centered on the observation ($Z_n \triangleq Z_r + jZ_i$) (as shown in Figure 5.15), the expectation of the product of the squared amplitude of the speech and noise E_s is calculated for each mixture j as

$$\begin{aligned} E_s^{(j)} &\triangleq \mathrm{E}\left(\left(R^{2(j)}_{n|n} + I^{2(j)}_{n|n}\right)\left(\left(R^{(j)}_{n|n} - Z_r\right)^2 + \left(I^{(j)}_{n|n} - Z_i\right)^2\right)\right) \\ &= \mathrm{E}\left(\left(R^{2(j)}_{n|n} + I^{2(j)}_{n|n}\right)^2\right) + \mathrm{E}\left(\left(Z_r^2 + Z_i^2\right)\left(R^{2(j)}_{n|n} + I^{2(j)}_{n|n}\right)\right) - 2\mathrm{E}\left(Z_r R^{3(j)}_{n|n} + Z_i I^{3(j)}_{n|n}\right) \\ &= \mu^{2(j)}_{ss} + \sigma^{2(j)}_{ss} + \left(Z_r^2 + Z_i^2\right)\mu^{(j)}_{ss} - 2Z_r\mathrm{E}\left(R^{3(j)}_{n|n}\right) - 2Z_i\mathrm{E}\left(I^{3(j)}_{n|n}\right) \end{aligned} \tag{5.61}$$

Since the real and imaginary parts, $R^{(j)}_{n|n}$ and $I^{(j)}_{n|n}$, are both Gaussian distributed, the expectation $\mathrm{E}\left(R^{3(j)}_{n|n}\right)$ and $\mathrm{E}\left(I^{3(j)}_{n|n}\right)$ can be calculated by making use of the moment generating function. Thus the correlation coefficient ρ_γ can be calculated as

$$\rho^{(j)}_\gamma = \frac{E^{(j)}_s - \widetilde{\Omega}^{(j)}\breve{\Omega}^{(j)}}{\sqrt{\widetilde{\sigma}^{2(j)}_{ss}\breve{\sigma}^{2(j)}_{ss}}} \tag{5.62}$$

where $\left(\widetilde{\Omega}, \tilde{\sigma}_{ss}^2\right)$ and $\left(\breve{\Omega}, \breve{\sigma}_{ss}^2\right)$ represent the parameters for speech and noise, respectively. The correlation coefficient of each mixture, $\rho_a^{(j)}$, can then be approximated by

$$\rho_a^{(j)} \approx \rho_\gamma^{(j)} \tag{5.63}$$

The covariance of the a posterior estimate of speech and noise amplitudes can now be calculated using statistics of the a posteriori estimate which have been derived above:

$$\begin{aligned}
\mathrm{E}\left(\widetilde{A}_{n|n}^{(j)}\breve{A}_{n|n}^{(j)}\right) &= \rho_a^{(j)}\sqrt{\mathrm{Var}\left(\widetilde{A}_{n|n}\right)\mathrm{Var}\left(\breve{A}_{n|n}\right)} + \mathrm{E}\left(\widetilde{A}_{n|n}^{(j)}\right)\mathrm{E}\left(\breve{A}_{n|n}^{(j)}\right) \\
&= \rho_a^{(j)}\sqrt{\mathrm{E}\left(\widetilde{A}_{n|n}^2\right) - \mathrm{E}\left(\widetilde{A}_{n|n}\right)^2}\sqrt{\mathrm{E}\left(\breve{A}_{n|n}^2\right) - \mathrm{E}\left(\breve{A}_{n|n}\right)^2} \qquad (5.64) \\
&= \rho_a^{(j)}\sqrt{\tilde{\mu}_{ss}^{2(j)} - \tilde{\mu}_a^{2(j)}}\sqrt{\breve{\mu}_{ss}^{2(j)} - \breve{\mu}_a^{2(j)}} + \tilde{\mu}_a^{(j)}\breve{\mu}_a^{(j)} \qquad (5.65)
\end{aligned}$$

$$v_a \triangleq \mathrm{E}\left(\widetilde{A}_{n|n}\breve{A}_{n|n}\right) - \mathrm{E}\left(\widetilde{A}_{n|n}\right)\mathrm{E}\left(\breve{A}_{n|n}\right) = \sum_{j=1}^{J_s J_n} \epsilon_{n|n}^{(j)}\left\{\mathrm{E}\left(\widetilde{A}_{n|n}^{(j)}\breve{A}_{n|n}^{(j)}\right)\right\} \tag{5.66}$$

The expression in (5.65) is obtained by substituting (5.53) into (5.64). Thus, the a posterior estimate of the speech and noise amplitudes have been obtained with the mean vector, $\boldsymbol{\mu}_{n|n} = \left[\tilde{\mu}_{n|n}\ \breve{\mu}_{n|n}\right]^T$ and the covariance matrix $\mathbf{V}_{n|n} = \begin{bmatrix} \tilde{\sigma}_{n|n}^2 & v_a \\ v_a & \breve{\sigma}_{n|n}^2 \end{bmatrix}$. In this section, the entire process of calculating the posterior estimate of both speech and noise from their prior estimate. has been given. First, the parameters of a Nakagami distribution are calculated by fitting to the prior estimate of speech and noise using (5.48) and (5.49) and get the parameters of the corresponding Rician distribution from them using (5.45) and (5.46). Thus, the mean and covariance of each Gaussian mixture component are obtained from (5.50) to (5.52) and the posterior distribution is obtained as the pairwise product of the components of speech and noise. Second, the parameters of the amplitude distribution for each

component of the posterior distribution are calculated using (5.55) and (5.56) and the overall mean and variance of the amplitude is obtained using (5.58) and (5.59). Third, the covariance of the amplitudes of speech and noise is approximated by that of the power of the amplitudes of speech and noise, which is given in (5.66).

5.3.2.3. Update of state vector

The final step is to update the entire state vector and the associated covariance matrix, $\mathbf{s}_{n|n}$ and $\mathbf{\Sigma}_{n|n}$. In this section a similar method is employed as in Section 5.2.5 and it is extended to the two-dimension case. Firstly, the prior state vector, $\mathbf{s}_{n|n-1}$, is permuted by swapping the second element and the $p+1$th element to make the first two elements corresponding to the speech and noise amplitudes, and the corresponding rows and columns in the covariance matrix $\mathbf{\Sigma}_{n|n-1}$. the covariance matrix $\mathbf{\Sigma}_{n|n-1}$ are then decomposed as

$$\mathbf{\Sigma}_{n|n-1} = \begin{bmatrix} \mathbf{V}_{n|n-1} & \mathbf{M}_n^T \\ \mathbf{M}_n & \mathbf{T}_n \end{bmatrix} \tag{5.67}$$

where $\mathbf{M}_n$ is a $(p+q-2) \times (p+q)$ matrix and $\mathbf{T}_n$ is a $(p+q-2) \times (p+q-2)$ matrix. the state vector are now transformed using the matrix

$$\mathbf{H}_n = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0}^T \\ -\mathbf{M}_n \mathbf{V}_{n|n-1}^{-1} & \mathbf{I}_{(p+q-2)} \end{bmatrix}$$

where $\mathbf{I}_2$ is a 2×2 identity matrix and $\mathbf{I}_{(p+q-2)}$ is the a $(p+q-2) \times (p+q-2)$ identity matrix.

$$\mathbf{z}_{n|n-1} = \mathbf{H}_n \mathbf{s}_{n|n-1} = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0}^T \\ -\mathbf{M}_n \mathbf{V}_{n|n-1}^{-1} & \mathbf{I}_{(p+q-2)} \end{bmatrix} \mathbf{s}_{n|n-1}$$

The autocorrelation matrix of $\mathbf{z}_{n|n-1}$ is given by

$$\begin{aligned}
\mathrm{E}\left(\mathbf{z}_{n|n-1}\mathbf{z}_{n|n-1}^T\right) &= \mathbf{H}_n \mathrm{E}\left(\mathbf{s}_{n|n-1}\mathbf{s}_{n|n-1}^T\right)\mathbf{H}_n^T = \mathbf{H}_n \boldsymbol{\Sigma}_{n|n-1}\mathbf{H}_n^T \\
&= \begin{bmatrix} \mathbf{I}_2 & \mathbf{0}^T \\ -\mathbf{M}_n\mathbf{V}_{n|n-1}^{-1} & \mathbf{I}_{(p+q-2)} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{n|n-1} & \mathbf{M}_n^T \\ \mathbf{M}_n^T & \mathbf{T}_n \end{bmatrix} \begin{bmatrix} \mathbf{I}_2 & -\mathbf{V}_{n|n-1}^{-1}\mathbf{M}_n^T \\ \mathbf{0} & \mathbf{I}_{(p+q-2)} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{V}_{n|n-1} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{T}_n - \mathbf{M}_n^T(\mathbf{V}_{n|n-1}^{-1})\mathbf{M}_n \end{bmatrix}
\end{aligned}$$

It can be seen that the vector consisting of first two elements in the augmented state vector is uncorrelated with other blocks and is distributed as $\mathcal{N}(\mathbf{C}^T\mathbf{z}_{n|n-1}, \mathbf{V}_{n|n-1})$ with $\mathbf{C} = \begin{bmatrix} \mathbf{I}_2 \\ \mathbf{0} \end{bmatrix}$, where $\mathbf{0}$ is a $(p+q-2)\times 2$ matrix of zeros.

the posterior distribution of the current sample,

$\mathcal{N}(\mathbf{C}^T\mathbf{z}_{n|n-1}; \boldsymbol{\mu}_{n|n-1}, \mathbf{V}_{\mathbf{n|n}})$, has been obtained, as a result, $\mathbf{C}^T\mathbf{z}_{n|n}$ can be obtained as

$$\mathbf{C}^T\mathbf{z}_{n|n} = \mathbf{C}^T\mathbf{z}_{n|n-1} + (\boldsymbol{\mu}_{n|n-1} - \mathbf{C}^T\mathbf{z}'_{n|n-1})$$

Thus for the entire transformed state vector, $\mathbf{z}_{n|n}$

$$\mathbf{z}_{n|n} = \mathbf{z}_{n|n-1} + \mathbf{C}\left(\boldsymbol{\mu}_{n|n-1} - \mathbf{C}^T\mathbf{z}_{n|n-1}\right) \tag{5.68}$$

$$\mathbf{s}_{n|n-1} = \mathbf{H}_n^{-1}\left(\mathbf{z}_{n|n-1} + \mathbf{C}\left(\boldsymbol{\mu}_{n|n-1} - \mathbf{C}^T\mathbf{z}_{n|n-1}\right)\right) \tag{5.69}$$

The covariance matrix, $\mathbf{\Sigma}_{n|n}$, can be calculated as

$$\begin{aligned}\mathbf{\Sigma}_{n|n} &= \mathbf{H}_n^{-1}\begin{bmatrix}\mathbf{V}_{n|n-1} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{T}_n - \mathbf{M}_n^T(\mathbf{C}_{n|n-1}^{-1})\mathbf{M}_n\end{bmatrix}(\mathbf{H}_n^{-1})^T \\ &= \begin{bmatrix}\mathbf{V}_{n|n-1} & \mathbf{V}_{n|n}(\mathbf{V}_{n|n-1}^{-1})\mathbf{M}_n^T \\ \mathbf{M}_n(\mathbf{V}_{n|n-1}^{-1})\mathbf{V}_{n|n} & \mathbf{T}_n - \mathbf{M}_n^T(\mathbf{V}_{n|n-1}^{-1})\mathbf{M}_n + \mathbf{M}_n(\mathbf{V}_{n|n-1}^{-1})\mathbf{V}_{n|n}(\mathbf{V}_{n|n-1}^{-1})\mathbf{M}_n^T\end{bmatrix} \\ &= \mathbf{\Sigma}_{n|n-1} - \mathbf{\Sigma}_{n|n-1}\mathbf{C}\mathbf{V}_{n|n-1}^{-1}\mathbf{C}^T\mathbf{\Sigma}_{n|n-1} + \mathbf{\Sigma}_{n|n-1}\mathbf{C}\mathbf{V}_{n|n-1}^{-1}\mathbf{V}_{n|n}\mathbf{V}_{n|n-1}^{-1}\mathbf{C}^T\mathbf{\Sigma}_{n|n-1}\end{aligned} \tag{5.70}$$

5.3.2.4. Implementation and evaluation

In this section, the performance of the proposed Modulation Domain Kalman filter based on a Gaussring model (MDKFR) are compared with logMMSE enhancer [53], the Modulation Domain Kalman filter that assumes colored noise (MDKFC) from [66] and the KMMSE algorithm introduced in Section 5.2, which, like MDKFR, extract the modulation-domain speech LPC coefficients from the enhanced speech using the logMMSE enhancer.

The parameters of all the algorithms were chosen similarly to those described in Section 3.2.4 where an acoustic frame length of 16 ms with a 4 ms increment is used which gives a 250 Hz sampling frequency in the modulation domain. The modulation frame length is 64 ms with a 16 ms increment. The speech LPC model is estimated from each modulation frame of the logMMSE enhanced speech and the noise model is estimated from the noise modulation power spectrum which is estimated using a VAD as in Section 2.5.1. The model orders for the speech and noise are $p = 3$ and $q = 4$ respectively. The sensitivity of the orders of LPC models of speech and noise has been discussed in Section 1.4.1.4 and 1.4.2.5. The test speech and noise are selected in the same vein as Section 5.2.7. The speech are corrupted

at $-10, -5$, 0, 5, 10 and 15 dB global SNR.

As in the previous chapters, the performances of the algorithms are evaluated using both segSNR and the PESQ measure. All the measured values shown are averages over all the sentences (totalling 192 sentences) in the TIMIT core test set. Figures 5.17 and 5.18 show the average segSNR for car noise and street noise, respectively. It can be seen that, for low SNRs, both MDKFR and MDKFC algorithms give similar performance, which outperform the logMMSE enhancer by about 5 dB for car noise and by about 4 dB for street noise. At high SNRs, the MDKFC algorithm performs better than the MDKFR algorithm, and the MDKFR algorithm give similar performance to the logMMSE enhancer at 15 dB SNR. The reason for this may be that, at high SNRs, the speech distortion introduced by the approximation of Gaussring is much more obvious, thus the MDKFR algorithm may give worse performance than MDKFC algorithm for segSNR.

Figures 5.19 and 5.20 give the corresponding average PESQ performance at each SNR. It shows that for car noise, all the enhancers give similar performance at 15 dB global SNR. At low SNRs, the MDKFR algorithm gives an improvement of about 0.1 and 0.25 PESQ over the MDKFC algorithm and logMMSE algorithm, respectively. The average PESQ values for street noise are given in Figure 5.20. It can be seen that the overall trend is the same as that for the car noise, except that at 15 dB the MDKFR and MDKFC algorithms also give better performance than that of the logMMSE enhancer, which is about 0.1 and 0.2 respectively. Similar to Figure 5.10, Figure 5.21 shows box plots of the difference in PESQ score between MDKFR and competing algorithms. It can be seen that in all cases the entire box lies below the axis line; this indicates that MDKFR results in an improvement for an overwhelming majority of speech-plus-noise combinations.

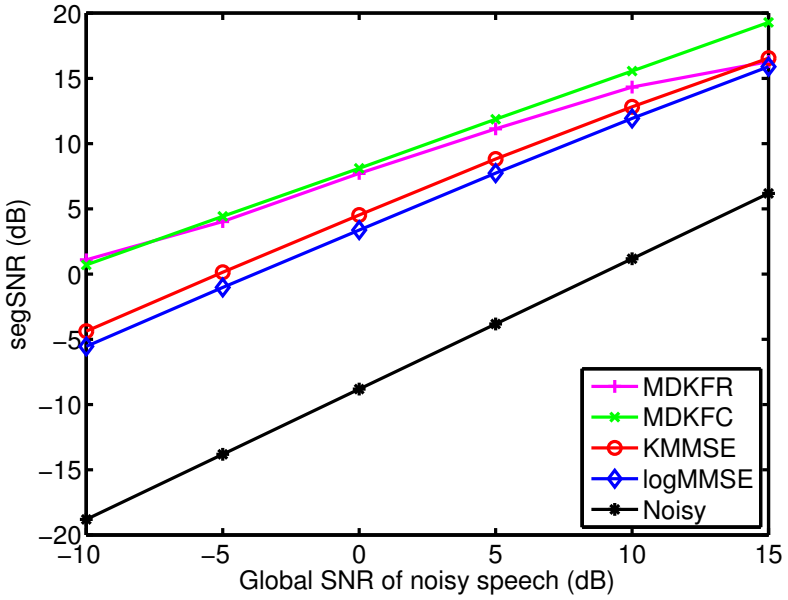

Figure 5.17.: Average segmental SNR of enhanced speech after processing by four algorithms plotted against the global SNR of the input speech corrupted by additive car noise. The algorithm acronyms are defined in the text.

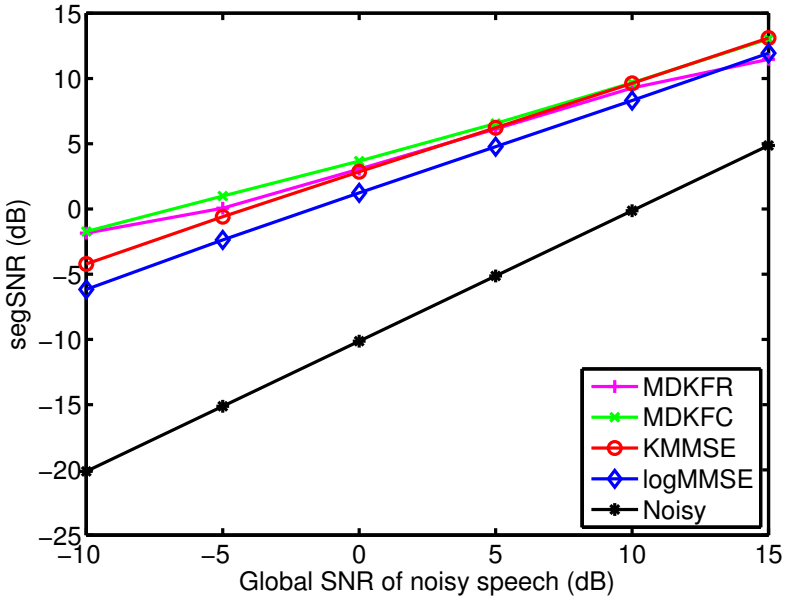

Figure 5.18.: Average segmental SNR of enhanced speech after processing by four algorithms plotted against the global SNR of the input speech corrupted by additive street noise.

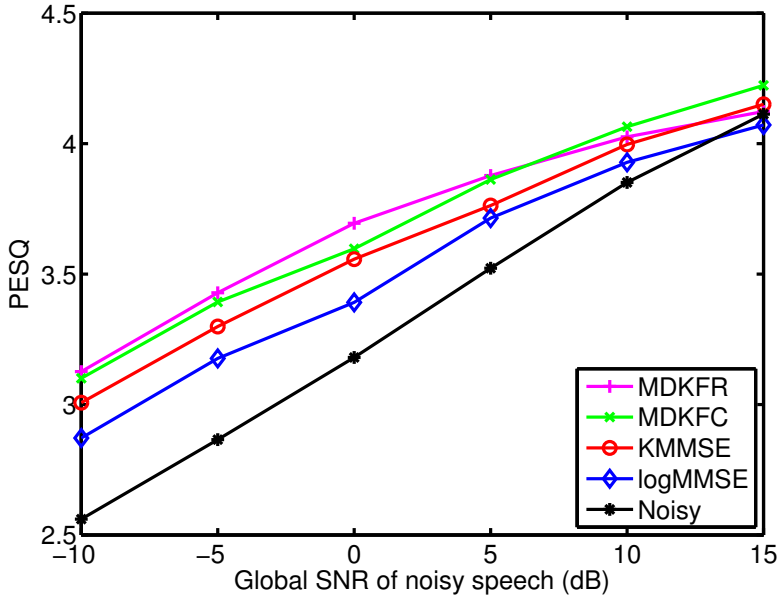

Figure 5.19.: Average PESQ of enhanced speech after processing by four algorithms plotted against the global SNR of the input speech corrupted by additive car noise.

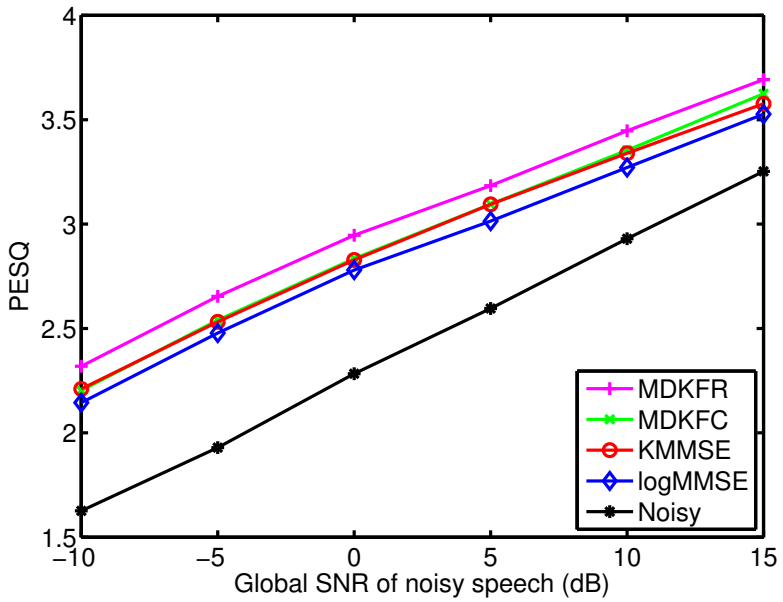

Figure 5.20.: Average PESQ of enhanced speech after processing by four algorithms plotted against the global SNR of the input speech corrupted by additive street noise.

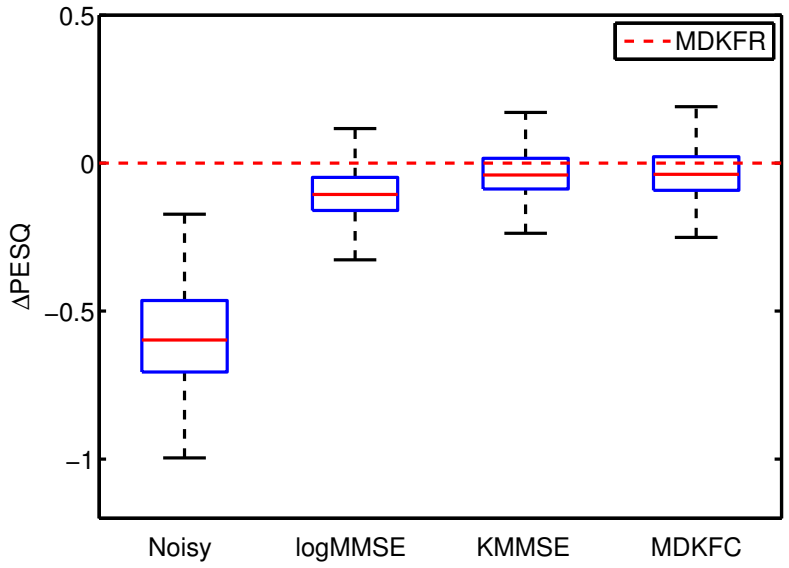

Figure 5.21.: Box plot showing the difference in PESQ score between competing algorithms and the proposed algorithm, MDKFR for 2376 speech+noise combinations.

5.3.3. Conclusion

In this chapter two different methods of estimating the spectral amplitudes of the clean speech based on a modulation Kalman filter have been proposed. The first methods incorporates the dynamics of spectral amplitudes of the speech, where the spectral amplitudes are assumed to follow a form of generalized Gamma distribution while the noise is assumed be Gaussian distributed. The second method also incorporates the spectral dynamics of the noise. To obtain the estimation of spectral amplitudes of both the speech and noise, a Gaussring model is proposed where mixtures of Gaussians are employed to model the prior distribution of the speech and noise in the complex Fourier domain, which leads to the proposed MDKFR algorithm. At low SNRs (< 5 dB), MDKFR gives consistently better PESQ results than all the competing algorithms. However at higher SNRs its performance degrades for reasons that may be related to the approximations made when using the Gaussring

prior.

6. Conclusions and Further Work

6.1. Summary of contributions

In this thesis we have investigated three approaches for implementing single-channel speech enhancement in the modulation domain. The goal in all three cases is to take advantage of prior knowledge about the temporal modulation of short-time spectral amplitudes. The first approach, described in Chapter 3, is to post-process the output of a conventional MMSE time-frequency domain enhancer using a Kalman filter in the modulation domain. The second approach, described in Chapter 4, performs enhancement directly in the modulation domain under the assumption that, within each frequency bin, the time-series of spectral amplitudes lies within a low dimensional subspace. Finally the third approach, described in Chapter 5, uses a modulation-domain Kalman filter to perform enhancement using two alternative distribution families for the speech and noise amplitude prior distributions.

6.1.1. Modulation domain post-processing

In Chapter 3, we have proposed two different methods of post-processing the output of an MMSE spectral amplitude speech enhancer by using a Kalman filter in the modulation domain. In the first part of the chapter, different modulation-domain

LPC models are introduced for speech and it shows that the post-processors based on the LPC models give consistent improvements over the MMSE enhancer, where the postprocessor using the original LPC model is denoted as KFMD algorithm. In the second part of the thesis, a post-processor in the modulation domain is introduced (KFGM) which uses a GMM for modeling prediction error of the noise in the output spectral amplitude of MMSE enhancer. The derivation of a Kalman filter that incorporates a GMM noise model has been given and a method for adaptively updating the GMM parameters during the processing has also been presented

6.1.2. Modulation domain subspace enhancement

In Chapter 4, a speech enhancement algorithm using subspace decomposition technique in the short-time modulation domain (MDSS) has been presented. A method to precompute the whitening matrix has also been proposed. The performance of the proposed enhancer has been evaluated using segSNR and PESQ and it shows that, for both stationary and non-stationary colored noise, it outperforms a time-domain subspace enhancer and a modulation-domain spectral-subtraction enhancer.

6.1.3. Modulation domain Kalman filtering

In Chapter 6, two different methods of estimating the spectral amplitudes of the clean speech based on a modulation Kalman filter have been proposed. In the first part of the chapter, an algorithm (KMMSE) which incorporates the dynamics of spectral amplitudes of the speech into the MMSE amplitudes estimation, in which the speech spectral amplitudes are assumed to follow a form of generalized Gamma distribution and the noise is assumed be Gaussian distributed. The second method (MDKFR) also incorporates the spectral dynamics of the noise by introducing a

novel prior distribution, Gaussring, in which mixtures of Gaussians are employed to model the prior distribution of the speech and noise in the complex Fourier domain.

6.2. Comparison of proposed algorithms

The three methods proposed in this thesis can be applied for different environmental conditions. For noisy speech of very low SNRs (< -15 dB), it is better to firstly pre-enhance the noisy speech and secondly post-process the initially enhanced speech using the CKFGM algorithm, which, as shown in Section 3.3.3, improves the segSNR by about 17 dB and the PESQ by about 0.5 over the noisy speech at −10 dB under non-stationary noise conditions. For noisy speech corrupted by stationary noise at low SNRs, the MDKFR algorithm should be used because under this circumstance, the modulation-domain LPC model of noise is easier to estimate which will bring great benefits together with the LPC model of the speech. However, the performance of the MDKFR algorithm degrades at high SNRs (> 10 dB) as shown in Section 5.3.2.4. For non-stationary noise conditions, it is better to use KMMSE or MDSS algorithms because the noise LPC model is difficult to estimate using the VAD-based method and only the acoustic-domain power of the noise should be made use of. Comparing to the MDSS algorithm, KMMSE algorithm may be better for colored noise because of the assumption made in the derivation of the noise covariance matrix in calculating the MDSS estimator. In Section 4.4.1 and Section 5.2.7, it can be seen that MDSS algorithm gives PESQ improvement of about 0.5 over noisy corrupted by factory noise at −10 dB, while the KMMSE algorithm gives PESQ improvement of about 0.6 for car noise and 0.8 for street noise at −10 dB. However, the MDSS estimator is more efficient to calculate because the autocorrelation sequence of the noise spectral amplitudes can be precomputed.

The Real-Time Factor (RTF) is introduced to compare the computational complexity of the proposed algorithms. The RTF is defined as the algorithm processing time divided by the duration of the speech signal for each algorithm determined by measuring the elapsed processing time using the Matlab cpu time function for each call on a 2.9 GHz Intel i5 Core processor with 8 GB 1.60 GHz DDR3 memory, and calculating the mean time per algorithm divided by the mean speech file duration. The RTFs of the proposed algorithms are listed in Table 6.1. It shows that KFMD, KFGM and KMMSE algorithms have similar computations while KFGM and MDKFR algorithms are much more computationally complex.

Algorithm	KFMD	KFGM	MDSS	KMMSE	MDKFR
RTF	4.11	13.42	4.03	4.02	12.92

Table 6.1.: RTFs of proposed algorithms.

6.3. Future Work

In this section, we describe a number of ways in which the work described in this thesis could be extended.

6.3.1. Better noise modulation power spectrum estimation

In order to estimate the modulation-domain LPC model of the noise, in this thesis we need to estimate its modulation power spectrum and the LPC coefficients of the noise amplitude sequence can be estimated from the power spectrum using Levinson-Durbin recursion. In Chapter 3 and Chapter 5 a SNR-based VAD is applied in the modulation domain to find the modulation frames where speech is absent. Since it is difficult to detect speech absence periods for non-stationary noise, better noise

power estimation methods, such as the Minimum Statistic method [41] and MMSE-based method [43], are worthwhile to explore in the modulation domain. In their current form, these noise power estimators attempt only to estimate the mean power spectrum of the noise. However, they could be extended to estimate a low-order LPC model for the noise instead.

6.3.2. Better LPC model

The LPC models applied in the modulation-domain have different characteristics from these of the time-domain LPC models. As described in Chapter 3, the time-domain LPC modes are used for modeling a sequence of time-domain signal with zero mean while in the modulation domain LPC model is applied for a sequence of positive spectral amplitudes whose mean is not zero. As mentioned in Section 1.4.1.4, in Kalman filter enhancers presented in this thesis, we have applied a positive floor to the amplitudes of speech and noise predicted by the LPC models. There may be better prediction models which give the optimal estimate of spectral amplitudes under the constraint that the predicted outputs are positive-valued signal.

6.3.3. Better Gaussring model

As shown in Section 5.3.2.4, the performance of the MDKFR algorithm degrades at high SNRs. Therefore, one future work is to understand why the performance of MDKFR algorithm degrades and to prevent this happening.

6.3.4. Incorporation of prior phase information

Most techniques for performing speech enhancement in the time-frequency domain preserve the phase spectrum of the noisy speech. The justification for this is that

the phase spectrum is less important perceptually than the amplitude spectrum and that, in the absence of prior knowledge, the phase spectrum of the noisy speech is the MMSE estimate of the phase spectrum of the clean speech. For voiced speech it is, however, possible to use knowledge of the pitch to obtain an improved estimate of the clean speech phase. It has been found that this can result in improved quality of enhanced speech [101]. It is straightforward to incorporate prior phase information in the Gaussring distribution described in Section 5.3.1 by applying different weights to the mixture components within the GMM.

6.3.5. Better domain for processing

It is also possible to transform the amplitude spectrum into an alternative domain (e.g. cepstrum) in which speech signals are sparser. Processing the signals in these domains could improve performance and possibly reduce computation by compressing the number of frequency channels.

A. Special Functions

A.1. Hypergeometric Function

A.1.1. Gauss Hypergeometric Function

All the descriptions provided in this subsection have been extracted from [89].

The Gauss hypergeometric function ${}_2F_1\left(a,b,c;z\right)$ is defined by the Gauss series

$$\begin{aligned} {}_2F_1\left(a,b,c;z\right) &= \sum_{s=0}^{\infty} \frac{\left(a\right)_s \left(b\right)_s}{\left(c\right)_s s!} z^s = 1 + \frac{ab}{c}z + \frac{a\left(a+1\right)b\left(b+1\right)}{c(c+1)2!} + \cdots \\ &= \frac{\Gamma\left(c\right)}{\Gamma\left(a\right)\Gamma\left(b\right)} \sum_{s=0}^{\infty} \frac{\Gamma\left(a+s\right)\Gamma\left(b+s\right)}{\Gamma\left(c+s\right)s!} z^s, \end{aligned} \tag{A.1}$$

on the disk $|z| < 1$, and by analytic continuation elsewhere. In general, ${}_2F_1\left(a,b,c;z\right)$ does not exist when $c = 0, -1, -2, \ldots$ The brach obtained by introducing a cut from 1 to $+\infty$ on the real z-axis, that is, the branch in the sector $|\mathrm{ph}(1-z)| \leq \pi$, where $\mathrm{ph}\left(\cdot\right)$ is the operator calculating the phase in the range $[-\pi, \pi]$, is the *principal branch* (or *principle value*) of ${}_2F_1\left(a,b,c;z\right)$.

A.1.2. Confluent Hypergeometric Function

Confluent hypergeometric differential equation is given by

$$z\frac{\mathrm{d}^2w}{\mathrm{d}z^2} + (b-z)\frac{\mathrm{d}w}{\mathrm{d}z} - aw = 0$$

with a regular singular point at $z = 0$ and an irregular singular point at $z = \infty$. This equation is also known as Kummer's equation. Kummer's confluent hypergeometric function is a solution to this equation which is given by [89]

$$\mathcal{M}(a,b;z) = 1 + \frac{a}{b}z + \frac{a\,(a+1)}{b\,(b+1)}\frac{z^2}{2!} + \ldots = \sum_{k=0}^{\infty}\frac{(a)_k}{(b)_k}\frac{z^k}{k!} = {}_1F_1(a,b;z),$$

where $(a)_k$ and $(b)_k$are Pochhammer symbols and $(a)_k$ is given by [89]

$$\begin{aligned}(a)_0 &= 1\\ (a)_k &= \frac{\Gamma\,(a+k)}{\Gamma\,(a)} = a\,(a+1)\,(a+2)\cdots(a+k-1)\,.\end{aligned}$$

Because the Kummer's equation is a second-order equation, there is another independent solution which is named Tricomi's confluent hypergeometric function and normally denoted as $U(a,b;z)$ [89]. This function is defined in terms of the Kummer's confluent hypergeometric function ,$\mathcal{M}(a,b;z)$, by

$$U(a,b;z) = \frac{\Gamma\,(1-b)}{\Gamma\,(a-b+1)}\mathcal{M}(a,b;z) + \frac{\Gamma\,(b-1)}{\Gamma\,(a)}z^{1-b}\mathcal{M}(a-b+1,2-b;z).$$

A.2. Parabolic Cylinder Function

Parabolic cylinder functions are solutions of the differential equation [89].

$$\frac{\mathrm{d}^2 w}{\mathrm{d}z^2} + \left(az^2 + bz + c\right) w = 0, \tag{A.2}$$

with three distinct standard forms

$$\frac{\mathrm{d}^2 w}{\mathrm{d}z^2} - \left(\frac{1}{4}z^2 + a\right) w = 0 \tag{A.3}$$

$$\frac{\mathrm{d}^2 w}{\mathrm{d}z^2} + \left(\frac{1}{4}z^2 - a\right) w = 0 \tag{A.4}$$

$$\frac{\mathrm{d}^2 w}{\mathrm{d}z^2} + \left(\nu + \frac{1}{2} - \frac{1}{4}z^2\right) w = 0 \tag{A.5}$$

Each of these equations is transformable into the others. The parabolic cylinder function that is used in this thesis, $\mathcal{D}_v\left(z\right)$, is the solution to (A.5). It can be defined in terms of the Tricomi's confluent hypergeometric function by [89]

$$\mathcal{D}_v\left(z\right) = U\left(-\frac{1}{2} - \nu, z\right)$$

B. Derivations

B.1. Derivations of MMSE Estimator in 5.18

In this section the details of the derivation of the MMSE estimator in 5.18 are given. The estimator is given by

$$\begin{aligned}\tilde{\mu}_{n|n} &= \int_0^\infty a_n p(a_n|\mathcal{R}_n) da_n = \frac{\int_0^\infty \int_0^{2\pi} a_n^{2\gamma_n} \exp\left\{-\frac{a_n^2}{\beta_n^2} - \frac{1}{\nu_n^2}|z_n - a_n e^{j\phi_n}|^2\right\} d\phi_n da_n}{\int_0^\infty \int_0^{2\pi} a_n^{2\gamma_n - 1} \exp\left\{-\frac{a_n^2}{\beta_n^2} - \frac{1}{\nu_n^2}|z_n - a_n e^{j\phi_n}|^2\right\} d\phi_n da_n} \\ &= \frac{\int_0^\infty \int_0^{2\pi} a_n^{2\gamma_n} \exp\left\{-\frac{a_n^2}{\beta_n^2} - \frac{1}{\nu_n^2}|z_n - a_n\left(\cos\phi_n + j\sin\phi_n\right)|^2\right\} d\phi_n da_n}{\int_0^\infty \int_0^{2\pi} a_n^{2\gamma_n - 1} \exp\left\{-\frac{a_n^2}{\beta_n^2} - \frac{1}{\nu_n^2}|z_n - a_n\left(\cos\phi_n + j\sin\phi_n\right)|^2\right\} d\phi_n da_n} \\ &= \frac{\int_0^\infty \int_0^{2\pi} a_n^{2\gamma_n} \exp\left\{-\frac{a_n^2}{\beta_n^2} - \frac{1}{\nu_n^2}\left(z_n^2 + a_n^2 - 2z_n a_n \cos\phi_n\right)\right\} d\phi_n da_n}{\int_0^\infty \int_0^{2\pi} a_n^{2\gamma_n - 1} \exp\left\{-\frac{a_n^2}{\beta_n^2} - \frac{1}{\nu_n^2}\left(z_n^2 + a_n^2 - 2z_n a_n \cos\phi_n\right)\right\} d\phi_n da_n}. \end{aligned} \tag{B.1}$$

By using the integral representation of the modified Bessel function of the nth order [89], which is defined as

$$I_n(x) = \frac{1}{2\pi} \int_0^{2\pi} \cos\alpha n \exp\left(x\cos\alpha\right) d\alpha, \tag{B.2}$$

(B.1) becomes

$$\tilde{\mu}_{n|n} = \frac{\int_0^\infty a_n^{2\gamma_n} \exp\left\{-\frac{a_n^2}{\beta_n^2} - \frac{a_n^2}{\nu_n^2}\right\} I_0\left(\frac{2a_n r_n}{\nu_n^2}\right) da_n}{\int_0^\infty a_n^{2\gamma_n - 1} \exp\left\{-\frac{a_n^2}{\beta_n^2} - \frac{a_n^2}{\nu_n^2}\right\} I_0\left(\frac{2a_n r_n}{\nu_n^2}\right) da_n}. \tag{B.3}$$

where r_n represents a realization of the random variable R_n. Defining $t \triangleq a_n^2$ $(t \geq 0)$ and substituting it in (B.3), we have

$$\tilde{\mu}_{n|n} = \frac{\int_0^\infty t^{\gamma_n - \frac{1}{2}} \exp\left\{-\frac{t}{\beta_n^2} - \frac{t}{\nu_n^2}\right\} I_0\left(\frac{r_n}{\nu_n^2}\sqrt{t}\right) dt}{\int_0^\infty t^{\gamma_n - 1} \exp\left\{-\frac{t}{\beta_n^2} - \frac{t}{\nu_n^2}\right\} I_0\left(\frac{r_n}{\nu_n^2}\sqrt{t}\right) dt}$$

By using [94, Eq. 6.643.2 and 9.220.2], (B.3) becomes a closed form equation

$$\tilde{\mu}_{n|n} = \frac{\Gamma\left(\gamma_n + 0.5\right)\sqrt{\left(\frac{\beta_n^2\nu_n^2}{\beta_n^2+\nu_n^2}\right)}\mathcal{M}\left(\gamma_n + 0.5; 1; \frac{r_n^2\beta_n^2}{\nu_n^2(\beta_n^2+\nu_n^2)}\right)}{\Gamma\left(\gamma_n\right)\sqrt{\left(\frac{\beta_n^2\nu_n^2}{\beta_n^2+\nu_n^2}\right)}\mathcal{M}\left(\gamma_n; 1; \frac{r_n^2\beta_n^2}{\nu_n^2(\beta_n^2+\nu_n^2)}\right)}. \tag{B.4}$$

Because the a priori SNR and a posteriori SNR are calculated as

$$\xi_n = \frac{\mathrm{E}(A_n^2|\mathcal{R}_{n-1})}{\nu_n^2} = \frac{\tilde{\mu}_{n|n-1}^2 + \tilde{\sigma}_{n|n-1}^2}{\nu_n^2},\ \zeta_n = \frac{r_n^2}{\nu_n^2} \tag{B.5}$$

where $\tilde{\mu}_{n|n-1}$and $\tilde{\sigma}_{n|n-1}^2$ are defined in (5.9) and (5.10) respectively. Thus the a priori SNR can be calculated as

$$\xi_n = \frac{\gamma_n \beta_n^2}{\nu_n^2} \tag{B.6}$$

By substituting the a prior SNR in (B.6) and a posteriori SNR in (B.5) into (B.4), the estimator, $\tilde{\mu}_{n|n}$, can be obtained as

$$\tilde{\mu}_{n|n} = \frac{\Gamma\left(\gamma_n + 0.5\right)}{\Gamma\left(\gamma_n\right)}\sqrt{\frac{\xi_n}{\zeta_n(\gamma_n + \xi_n)}}\frac{\mathcal{M}\left(\gamma_n + 0.5; 1; \frac{\zeta_n\xi_n}{\gamma_n+\xi_n}\right)}{\mathcal{M}\left(\gamma_n; 1; \frac{\zeta_n\xi_n}{\gamma_n+\xi_n}\right)} r_n$$

B.2. Derivations of noise spectral amplitudes autocorrelation

In this section the derivation of noise spectral amplitudes autocorrelation given in (4.10) is given. The main steps of the derivation follow those of [88, Eq. 4.14].

Suppose $\mathbf{z} = \left[\widetilde{W}_{n,k}, \widetilde{W}_{n+\tau,k}\right]^T$ is a two dimensional complex Gaussian vector, where

$$\widetilde{W}_{n,k} = \left|\widetilde{W}_{n,k}\right| \exp\left(j\psi_1\right) \quad \widetilde{W}_{n+\tau,k} = \left|\widetilde{W}_{n+\tau,k}\right| \exp\left(j\psi_2\right)$$

and $\mathbf{z}$ has zero mean and positive definite covariance matrix $\mathbf{\Sigma}$. In this section, It is more convenient to use the precision matrix $\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^* & P_{22} \end{bmatrix}$ where $*$ represents complex conjugate. The elements in $\mathbf{P}$ are defined by

$$P_{11} = \frac{\mathrm{E}\left(\left|\widetilde{W}_{n+\tau,k}\right|^2\right)}{\mathrm{E}\left(\left|\widetilde{W}_{n,k}\right|^2\right)\mathrm{E}\left(\left|\widetilde{W}_{n+\tau,k}\right|^2\right) - \mathrm{E}\left(\widetilde{W}_{n,k}\widetilde{W}_{n+t,k}^*\right)\left(\mathrm{E}\left(\widetilde{W}_{n,k}\widetilde{W}_{n+t,k}^*\right)\right)^*}$$

$$P_{22} = \frac{\mathrm{E}\left(\left|\widetilde{W}_{n,k}\right|^2\right)}{\mathrm{E}\left(\left|\widetilde{W}_{n,k}\right|^2\right)\mathrm{E}\left(\left|\widetilde{W}_{n+\tau,k}\right|^2\right) - \mathrm{E}\left(\widetilde{W}_{n,k}\widetilde{W}_{n+t,k}^*\right)\left(\mathrm{E}\left(\widetilde{W}_{n,k}\widetilde{W}_{n+t,k}^*\right)\right)^*}$$

$$P_{12} = \frac{-\mathrm{E}\left(\widetilde{W}_{n,k}\widetilde{W}_{n+t,k}^*\right)}{\mathrm{E}\left(\left|\widetilde{W}_{n,k}\right|^2\right)\mathrm{E}\left(\left|\widetilde{W}_{n+\tau,k}\right|^2\right) - \mathrm{E}\left(\widetilde{W}_{n,k}\widetilde{W}_{n+t,k}^*\right)\left(\mathrm{E}\left(\widetilde{W}_{n,k}\widetilde{W}_{n+t,k}^*\right)\right)^*}$$

The expectations$\mathrm{E}\left(|\widetilde{W}_{n,k}|^2\right) = \mathrm{E}\left(|\widetilde{W}_{n+\tau,k}|^2\right)$ have been given in (4.8) and the expectation $\mathrm{E}\left(\widetilde{W}_{n,k}\widetilde{W}_{n+t,k}^*\right)$ has been obtained as (4.9), substituting them into the elements of $\mathbf{P}$ can obtain

$$P_{11} = P_{22} = \frac{1}{\nu_w^2 \left(1 - |\rho_h(\tau, k)|^2\right)} \tag{B.7}$$

$$P_{12} = \frac{-\rho_h(\tau, k)}{\nu_w^2 \left(1 - |\rho_h(\tau, k)|^2\right)} \tag{B.8}$$

From [88, Eq. 3.11], the joint distribution of the amplitude $\left|\widetilde{W}_{n,k}\right|$ and $\left|\widetilde{W}_{n+\tau,k}\right|$ is given by

$$p\left(|\widetilde{w}_{n,k}|\,|\widetilde{w}_{n+\tau,k}|\right) = 4\,|\widetilde{w}_{n,k}|\,|\widetilde{w}_{n+\tau,k}|\,(P_{11}+P_{22})\exp\left\{-\left(P_{11}\,|\widetilde{w}_{n,k}|^2 + P_{22}\,|\widetilde{w}_{n+\tau,k}|^2\right)\right\} \times I_0\left(2\,|P_{12}|\,|\widetilde{w}_{n,k}|\,|\widetilde{w}_{n+\tau,k}|\right) \tag{B.9}$$

where $\tilde{w}_{n,k}$ and $\tilde{w}_{n+\tau,k}$ represents a realization of $\widetilde{W}_{n,k}$ and $\widetilde{W}_{n+\tau,k}$, and I_0 is the Bessel function that is given in (B.2). The noise spectral amplitudes autocorrelation, $\mathrm{E}\left(\left|\widetilde{W}_{n,k}\right|\left|\widetilde{W}_{n+\tau,k}\right|\right)$, is given by

$$\mathrm{E}\left(\left|\widetilde{W}_{n,k}\right|\left|\widetilde{W}_{n+\tau,k}\right|\right) = \int_0^\infty \int_0^\infty |\widetilde{w}_{n,k}|\,|\widetilde{w}_{n+\tau,k}|\,p\left(|\widetilde{w}_{n,k}|\,|\widetilde{w}_{n+\tau,k}|\right) d\,|\widetilde{w}_{n,k}|\,d\,|\widetilde{w}_{n+\tau,k}| \tag{B.10}$$

Substituting (B.7), (B.8) and (B.9) into (B.10), it becomes

$$\mathrm{E}\left(\left|\widetilde{W}_{n,k}\right|\left|\widetilde{W}_{n+\tau,k}\right|\right) = \int_0^\infty \int_0^\infty \frac{8\,|\widetilde{w}_{n,k}|^2\,|\widetilde{w}_{n+\tau,k}|^2}{\nu_w^2\left(1-|\rho_h(\tau,k)|^2\right)} \exp\left\{-\left(\frac{|\widetilde{w}_{n,k}|^2 + |\widetilde{w}_{n+\tau,k}|^2}{\nu_w^2\left(1-|\rho_h(\tau,k)|^2\right)}\right)\right\} \times I_0\left(\frac{2\,|\rho_h(\tau,k)|}{\nu_w^2\left(1-|\rho_h(\tau,k)|^2\right)}\,|\widetilde{w}_{n,k}|\,|\widetilde{w}_{n+\tau,k}|\right) d\,|\widetilde{w}_{n,k}|\,d\,|\widetilde{w}_{n+\tau,k}|$$

according to [88, Eq. 4.19], the closed-form equation for this integral is given by

$$\mathrm{E}\left(\left|\widetilde{W}_{n,k}\right|\left|\widetilde{W}_{n+\tau,k}\right|\right) = \frac{\pi}{4}\nu_w^2\left(1-|\rho_h(\tau,k)|^2\right)^2 \times {}_2F_1\left(\frac{3}{2},\frac{3}{2},1;|\rho_h(\tau,k)|^2\right) \tag{B.11}$$

where ${}_2F_1(\cdot)$ is the Gauss hypergeometric function introduced in (A.1.1). Because $\left(1-|\rho_h(\tau,k)|^2\right)^2 {}_2F_1\left(\frac{3}{2},\frac{3}{2},1;|\rho_h(\tau,k)|^2\right) = {}_2F_1\left(-\frac{1}{2},-\frac{1}{2},1;|\rho_h(\tau,k)|^2\right)$ [89], (B.11) can be simplified to

$$\mathrm{E}\left(\left|\widetilde{W}_{n,k}\right|\left|\widetilde{W}_{n+\tau,k}\right|\right) = \frac{\pi}{4}\nu_w^2 \times {}_2F_1\left(-\frac{1}{2},-\frac{1}{2},1;|\rho_h(\tau,k)|^2\right) \tag{B.12}$$

Bibliography

[1] P. C. Loizou. *Speech Enhancement Theory and Practice*. Taylor & Francis, 2007.

[2] J. Benesty, J. Chen, and Y. Huang. *Microphone Array Signal Processing*. Springer-Verlag, Berlin, Germany, 2008.

[3] Y. Ephraim, H. Lev-Ari, and W. J. J. Roberts. A brief survey of speech enhancement. In *The Electronic Handbook*. CRC Press, second edition, February 2005.

[4] S. Haykin. *Adaptive Filter Theory*. Prentice-Hall, fourth edition, 2002.

[5] B. Widrow, J. R. Glover, Jr, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, Eugene Dong, Jr, and R. C. Goodlin. Adaptive noise cancelling: Principles and applications. *Proc. IEEE*, 63(12):1692–1716, 1975.

[6] D. Wang and J. S. Lim. The unimportance of phase in speech enhancement. *IEEE Trans. Acoust., Speech, Signal Process.*, 30(4):679–681, 1982.

[7] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Process.*, 32(6):1109–1121, December 1984.

[8] P. J. Wolfe and S. J. Godsill. Simple alternatives to the Ephraim and Malah

suppression rule for speech enhancement. In *Proc. IEEE Signal Processing Workshop on Statistical Signal Processing*, pages 496–499, August 2001.

[9] J. B. Allen. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Trans. Acoust., Speech, Signal Process.*, 25(3):235–238, June 1977.

[10] J. Allen and L. Radiner. A unified approach to short-time Fourier analysis and synthesis. *Proc. IEEE*, 65(11):1558–1564, 1977.

[11] J. S. Garofolo. Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database. Technical report, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, December 1988.

[12] O. Cappe. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Trans. Speech Audio Process.*, 2(2):345–349, April 1994.

[13] L. Atlas and S. A. Shamma. Joint acoustic and modulation frequency. *EURASIP Journal on Applied Signal Processing*, 7:668–675, June 2003.

[14] B. E. D. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech communication*, 25(1):117–132, August 1998.

[15] J. Tchorz and B. Kollmeier. Estimation of the signal-to-noise ratio with amplitude modulation spectrograms. *Speech Commun.*, 38:1–17, September 2002.

[16] H. J. M. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. *J. Acoust. Soc. Am.*, 67(1):318–326, January 1980.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for

intelligibility prediction of time frequency weighted noisy speech. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(7):2125–2136, September 2011.

[18] R. Drullman, J. M. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.*, 95(5):2670–2680, May 1994.

[19] R. Drullman, J. M. Festen, and R. Plomp. Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.*, 95(2):1053–1064, February 1994.

[20] D. Byrne, H. Dillon, K. Tran, S. Arlinger, K. Wilbraham, R. Cox, B. Hayerman, R. Hetu, J. Kei, C. Lui, J. Kiessling, M. N. Kotby, N. H. A. Nasser, W. A. H. El Kholy, Y. Nakanishi, H. Oyer, R. Powell, D. Stephens, , T. Sirimanna, G. Tavartkiladze, G. I. Frolenkov, S. Westerman, and C. Ludvigsen. An international comparison of long-term average speech spectra. *J. Acoust. Soc. Am.*, 96(4):2108–2120, October 1994.

[21] N. D. Gaubitch, M. Brookes, and P. A. Naylor. Blind channel identification in speech using the long-term average speech spectrum. In *Proc. European Signal Processing Conf. (EUSIPCO)*, Glasgow, August 2009.

[22] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, April 1975.

[23] Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Process.*, 3(4):251 –266, July 1995.

[24] T. Esch and P. Vary. Speech enhancement using a modified Kalman filter based on complex linear prediction and supergaussian priors. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4877–4880, April 2008.

[25] H. J. M. Steeneken and F. W. M. Geurtsen. Description of the RSG.10 noise data-base. Technical Report IZF 1988–3, TNO Institute for perception, 1988.

[26] Y. Hu and P. C. Loizou. A subspace approach for enhancing speech corrupted by colored noise. *IEEE Signal Process. Lett.*, 9(7):204–206, July 2002.

[27] Y. Ephraim and I. Cohen. Recent advancements in speech enhancement. In R. C. Dorf, editor, *The Electrical Engineering Handbook, Circuits, Signals, and Speech and Image Processing*. CRC Press, third edition, 2006.

[28] J. Benesty, S. Makino, and J. Chen, editors. *Speech Enhancement*. Springer, 2005.

[29] N. Virag. Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Trans. Speech Audio Process.*, 7(2):126 –137, March 1999.

[30] J. Sohn, N. S. Kim, and W. Sung. A statistical model-based voice activity detection. *IEEE Signal Process. Lett.*, 6(1):1–3, January 1999.

[31] J. Roberts. Modification to piecewise LPC-10E. Technical Report WP-21752, MITRE, 1978.

[32] R. McAulay and M. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust., Speech, Signal Process.*, 28(2):137–145, April 1980.

[33] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit. ITU-T recommendation G.729 annex b: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. *IEEE Communications Magazine*, 35(9):64–73, 1997.

[34] S. G. Tanyer and H. Ozer. Voice activity detection in nonstationary noise. *IEEE Trans. Speech Audio Process.*, 8(4):478–482, July 2000.

[35] D. Malah, R. V. Cox, and A. J. Accardi. Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 789–792, March 1999.

[36] I. Cohen and B. Berdugo. Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Process. Lett.*, 9(1):12–15, January 2002.

[37] I. Cohen. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans. Speech Audio Process.*, 11(5):466–475, September 2003.

[38] S. Rangachari, P. C. Loizou, and Y. Hu. A noise estimation algorithm with rapid adaptation for highly nonstationary environments. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 305–308, May 2004.

[39] S. Rangachari and P. C. Loizou. A noise-estimation algorithm for highly non-stationary environments. *Speech Communication*, 48(2):220–231, February 2006.

[40] R. Martin. Spectral subtraction based on minimum statistics. In *Proc. European Signal Processing Conf*, pages 1182–1185, September 1994.

[41] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.*, 9(5):504–512, July 2001.

[42] R. Martin. Bias compensation methods for minimum statistics noise power spectral density estimation. *Signal Processing*, 86(6):1215–1229, June 2006.

[43] T. Gerkmann and R. C. Hendriks. Unbiased MMSE-based noise power es-

timation with low complexity and low tracking delay. *IEEE Trans. Audio, Speech, Lang. Process.*, 20(4):1383–1393, May 2012.

[44] R. C. Hendriks, R. Heusdens, and J. Jensen. MMSE based noise PSD tracking with low complexity. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4266–4269, March 2010.

[45] B. Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26(1):17–32, February 1981.

[46] Y. Hu and P. C. Loizou. A generalized subspace approach for enhancing speech corrupted by colored noise. *IEEE Trans. Speech Audio Process.*, 11(4):334–341, July 2003.

[47] U. Mittal and N. Phamdo. Signal/noise KLT based approach for enhancing speech degraded by colored noise. *IEEE Trans. Speech Audio Process.*, 8(2):159–167, March 2000.

[48] H. Lev-Ari and Y. Ephraim. Extension of the signal subspace speech enhancement approach to colored noise. *IEEE Signal Process. Lett.*, 10(4):104–106, April 2003.

[49] Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Process.*, 3(4):251–266, July 1995.

[50] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Process.*, 27(2):113 – 120, April 1979.

[51] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 208–211, April 1979.

[52] P. Lockwood and J. Boudy. Experiments with a nonlinear spectral subtractor

(NSS), hidden Markov models and the projection, for robust recognition in cars. *Speech Communication*, 11:215–228, June 1992.

[53] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Process.*, 33(2):443–445, April 1985.

[54] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Process.*, 13(5):845–856, September 2005.

[55] T. Lotter and P. Vary. Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. *EURASIP Journal on Applied Signal Processing*, pages 1110–1126, January 2005.

[56] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. *IEEE Trans. Speech Audio Process.*, 15(6):1741–1752, August 2007.

[57] T. Lotter, C. Benien, and P. Vary. Multichannel speech enhancement using Bayesian spectral amplitude estimation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, April 2003.

[58] P. C. Loizou. Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum. *IEEE Trans. Speech Audio Process.*, 13(5):857–869, August 2005.

[59] P. J. Wolfe and S. J. Godsill. Towards a perceptually optimal spectral amplitude estimator for audio signal enhancement. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages II821–II824 vol.2, June 2000.

[60] A. H. Gray, Jr. and J. D. Markel. Distance measures for speech processing. *IEEE Trans. Acoust., Speech, Signal Process.*, 24(5):380–391, October 1976.

[61] C. H. You, S. N. Koh, and S. Rahardja. β-order MMSE spectral amplitude estimation for speech enhancement. *IEEE Trans. Speech Audio Process.*, 13(4):475–486, June 2005.

[62] E. Plourde and B. Champagne. Auditory-based spectral amplitude estimators for speech enhancement. *IEEE Trans. Speech Audio Process.*, 16(8):1614–1623, Nov 2008.

[63] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Trans. Speech Audio Process.*, 2(4):578–589, October 1994.

[64] R. Singh and P. Rao. Spectral subtraction speech enhancement with RASTA filtering. *Proc. of National Conference on Communications (NCC)*, 2007.

[65] K. Paliwal, K. Wojcicki, and B. Schwerin. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. *Speech Communication*, 52(5):450–475, 2010. The Matlab software is available online at URL: http://maxwell.me.gu.edu.au/spl/research/modspecsub/.

[66] S. So and K. Paliwal. Modulation-domain Kalman filtering for single-channel speech enhancement. *Speech Communication*, 53(6):818–829, July 2011.

[67] K. Paliwal, B. Schwerin, and K. Wójcicki. Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. *Speech Commun.*, 54:282–305, February 2012.

[68] J. D. Gibson, B. Koo, and S. D. Gray. Filtering of colored noise for speech enhancement and coding. *IEEE Trans. Signal Process.*, 39(8):1732–1742, August 1991.

[69] A. Rix, J. Beerends, M. Hollier, and A. Hekstra. Perceptual evaluation of

speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 749–752, May 2001.

[70] Z. Goh, K.-C. Tan, and B. T. G. Tan. Postprocessing method for suppressing musical noise generated by spectral subtraction. *IEEE Trans. Speech Audio Process.*, 6(3):287–292, May 1998.

[71] M. Klein and P. Kabal. Signal subspace speech enhancement with perceptual post-filtering. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages I–537–I–540, May 2002.

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

[73] T. Esch and P. Vary. Efficient musical noise suppression for speech enhancement system. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4409–4412, April 2009.

[74] R. Martin and C. Breithaupt. Speech enhancement in the DFT domain using Laplacian speech priors. In *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, pages 87–90, September 2003.

[75] Y. Wang and M. Brookes. Speech enhancement using a robust Kalman filter post-processing in the modulation domain. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7457–7461, May 2013.

[76] ITU-T P.830. Subjective performance evaluation of telephone band and wideband codecs, February 1996.

[77] T. H. Falk and Wai-Yip Chan. Nonintrusive speech quality estimation using

Gaussian mixture models. *IEEE Signal Processing Letters*, 13(2):108–111, February 2006.

[78] J. Hansen and B. Pellom. An effective quality evaluation protocol for speech enhancement algorithms. In *Proc. Intl. Conf. on Spoken Lang. Processing (ICSLP)*, volume 7, pages 2819–2822, December 1998.

[79] Y. Hu and P. C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Speech Audio Process.*, 16(1):229–238, January 2008.

[80] A. W. Rix, J. G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza. Objective assessment of speech and audio quality - technology and applications. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(6):1890–1901, November 2006.

[81] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl. Perceptual objective listening quality assessment (POLQA), the Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I-Temporal Alignment. *Journal of the Audio Engineering Society*, 61(6):366–384, June 2013.

[82] J. Makhoul. Linear prediction: A tutorial review. *Proc. IEEE*, 63(4):561–580, April 1975.

[83] M. Brookes. The matrix reference manual. `http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/intro.html`, 1998-2015.

[84] P. Kabal. Ill-conditioning and bandwidth expansion in linear prediction of speech. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages I–824 – I–827, April 2003.

[85] M. Brookes. VOICEBOX: A speech processing toolbox for MATLAB. `http://`

`www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html`, 1998-2014.

[86] D. A. Reynolds, T. F. Quatieri, and R. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1–3):19–41, January 2000.

[87] ITU-T P.501. Test signals for use in telephonometry, August 1996.

[88] K. S. Miller. *Complex stochastic processes: an introduction to theory and application*. Addison-Wesley Publishing Company, Advanced Book Program, 1974.

[89] F. Olver, D. Lozier, R. F. Boiszert, and C. W. Clark, editors. *NIST Handbook of Mathematical Functions: Companion to the Digital Library of Mathematical Functions*. Cambridge University Press, 2010. URL: http://dlmf.nist.gov/13.

[90] Y. Avargel and I. Cohen. On multiplicative transfer function approximation in the short-time Fourier transform domain. *IEEE Signal Process. Lett.*, 14(5):337–340, May 2007.

[91] B. Chen and P. C. Loizou. Speech enhancement using a MMSE short time spectral amplitude estimator with Laplacian speech modeling. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 1097–1100, March 2005.

[92] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.*, 9:504–512, July 2001.

[93] L. Norman, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. Wiley, 1994.

[94] A. Jeffrey and D. Zwillinger. *Table of Integrals, Series, and Products*. Academic Press, 6th edition, 2000.

[95] M. Nakagami. *The m-distribution – A general formula of intensity distribution of rapid fading*. Pergamon Press, 1960.

[96] L. C. Wang and C. T. Lea. Co-channel interference analysis of shadowed Rician channels. *IEEE Communications Letters*, 2(3):67–69, March 1998.

[97] D. Xie and W. Zhang. Estimating speech spectral amplitude based on the Nakagami approximation. *IEEE Signal Processing Letters*, 21(11):1375–1379, Nov 2014.

[98] P. J. Crepeau. Uncoded and coded performance of MFSK and DPSK in nakagami fading channels. *IEEE Transactions on Communications*, 40(3):487–493, March 1992.

[99] J. F. Paris. Nakagami-q (Hoyt) distribution function with applications. *Electronics Letters*, 45(4):210–211, February 2009.

[100] Z. Song, K. Zhang, L. Guan, and Y. Liang. Generating correlated Nakagami fading signals with arbitrary correlation and fading parameters. In *Proc. Intl. Conf. Commun. (ICC)*, volume 3, pages 1363–1367 vol.3, April 2002.

[101] T. Gerkmann and M. Krawczyk. MMSE-optimal spectral amplitude estimation given the STFT-phase. *IEEE Signal Processing Letters*, 20(2):129–132, Feb 2013.