Abstract

Noise reduction aims to improve the quality of noisy speech by suppressing the background noise in the signal. However, there is always a tradeoff between noise reduction and signal distortion—more noise reduction is always accompanied by more signal distortion. An evaluation of the intelligibility of speech processed by several noise reduction algorithms [23] showed that most noise reduction algorithms were not successful in improving the intelligibility of noisy speech.

In this thesis, we aim to incorporate the information from acoustic-phonetics to enhance the intelligibility of noise-reduced speech. Acoustic-phonetics studies the characteristics of speech and the acoustic cues that are important for speech intelligibility. We considered the following questions: what is the noise reduction algorithm that we should use, what are the acoustic cues that should be targeted, and how to incorporate this information into the design of the noise reduction system.

A Bayesian noise reduction method similar to the one proposed by Ephraim and Malah in [16] is employed. We first evaluate the goodness-of-fit of several parametric PDF models to the empirical speech data. For classified speech, we find that the Rayleigh and Gamma with a fixed shape parameter, $v = 5$, model the speech spectral amplitude equally well. The Gamma-MAP and Gamma-MMSE estimators are derived. The subjective and objective performances of these estimators are then
We also propose to apply a class-based cue-enhancement processing, similar to those performed in [21]. The processing directly manipulates the acoustic cues known to be important for speech intelligibility. We assume that the system has the sound class information of the input speech. The scheme aims to enhance the interclass and intraclass distinction of speech sounds. The intelligibility of speech processed by the proposed system is then compared to the intelligibility of speech processed by the Rayleigh-MMSE estimator [16].

The intelligibility evaluation shows that the proposed scheme enhances the detection of plosive and fricative sounds. However, it does not help in the intraclass discrimination of plosive sounds, and more tests need to be done to evaluate whether intraclass discrimination of fricatives is improved. The proposed scheme deteriorates the detection of nasal and affricate sounds.
Acknowledgments

I would like to express my gratitude to those individuals who have helped and supported me throughout the course of my study. First, and foremost, I thank my supervisor, Dr. Saeed Gazor, for his help, advice, encouragement, support, and most of all, kindness.

I would also like to thank my thesis examiners, Dr. Evelyn Morin and Dr. Ingrid Johnsrude, for their efforts in carefully reviewing this thesis and providing useful comments and suggestions. Thanks are also due to the volunteers who participated in the listening experiments. I truly appreciate their willingness to help. I would also like to express my gratitude to my supervisor, OGS, and NSERC for the financial support given throughout the course of my study.

Many thanks also to my colleagues in the AMSPL lab. To Saeed Akhavan, thank you for your help in proofreading this thesis and providing many valuable inputs. Also to Khalid Al-Suhaili and Ruhallah Alihemati, thank you for your friendship. It is a great privilege to know you. I will miss our group discussions. To all my friends in Kingston, who have made my time in this city to be a memorable one, thank you very much.

Special thanks go to my parents and Lenko Grigorov, for their love, care, support, understanding, and everything else. I feel very fortunate to have you.
List of Abbreviations

iid independent and identically distributed

LP Linear Prediction

MAP Maximum A-Posteriori

MMSE Minimum Mean Squared Error

MLE Maximum Likelihood Estimate

PDF Probability Distribution Function

STFT Short-Time Fourier Transform

STSA Short-Time Spectral Amplitude

KLT Karhunen-Loeve Transform

VAD Voice Activity Detector
# Table of Contents

Abstract i  

Acknowledgments iii  

List of Abbreviations iv  

Table of Contents v  

List of Tables viii  

List of Figures ix  

1 Introduction 1  

1.1 Background and Motivation 1  

1.2 Thesis Objectives 4  

1.3 Thesis Contributions 5  

1.4 Thesis Outline 5  

Chapter 2:  

    Background and Proposed System Overview 7  

2.1 Overview of Speech Perception 7  

2.2 Acoustic Characteristics of Speech Sounds 10  

v
Chapter 6:

Conclusions .......................................................... 89

6.1 Summary .......................................................... 89

6.2 Future Work ..................................................... 92

Bibliography .......................................................... 94
## List of Tables

2.1 Phonemes in American English [42, Ch.2] ........................................... 12

4.1 Equivalent SNR from Wang and Lim’s experiment ............................... 50

4.2 PDF assumptions and distortion measures used in Bayesian speech
    estimators proposed in the literature ............................................. 54

4.3 $\alpha_{\xi}$ and $\alpha_{W}$ used in the subjective quality evaluation. ............. 70

5.1 The phonetically-balanced word list used in the intelligibility test. ... 81

5.2 Total number of correct word recognition obtained from the intelligibility
    test ................................................................. 84

5.3 Interclass confusion matrices obtained from the intelligibility test. 85

5.4 Plosive intraclass confusion matrices obtained from the intelligibility
    test. ......................................................... 86

5.5 Fricative intraclass confusion matrices obtained from the intelligibility
    test ......................................................... 87
List of Figures

2.1 Waveform, spectrogram, and spectrum of vowel /i/ (eve). .............. 13
2.2 Waveforms of unvoiced and voiced plosives /k/ and /g/. .............. 15
2.3 Waveform, spectrogram, and spectrum plots of unvoiced and voiced fricatives /s/ and /z/. ............................................. 17
2.4 Waveform, spectrogram, and spectrum plots of nasals /m/ and /n/. 19
2.5 Block diagram of the STFT-domain noise reduction algorithm. .... 20
2.6 Summary of the class-based cue-enhancement processing. ............ 22

3.1 Rayleigh, Exponential, and Gamma PDFs .............................. 28
3.2 Average $v_k$ (shape parameter of the Gamma PDF) over all unclassified segments for segment lengths $N_f = 5, 10, 20, 30$. .......... 31
3.3 Percentages of unclassified segments with maximum $\mathcal{L}^{Rayl}_{X_k}$, $\mathcal{L}^{Exp}_{X_k}$, and $\mathcal{L}^{\gamma}_{X_k}$ values. ................................................. 33
3.3 [cont.] Percentages of unclassified segments with maximum $\mathcal{L}^{Rayl}_{X_k}$, $\mathcal{L}^{Exp}_{X_k}$, and $\mathcal{L}^{\gamma}_{X_k}$ values. ................................................. 34
3.4 Percentages of unclassified segments with minimum $\chi^2_{Rayl, X_k}$, $\chi^2_{Exp, X_k}$, and $\chi^2_{\gamma, X_k}$ values. ................................................. 35
3.5 Average $v_k$ (shape parameter of the Gamma PDF) over the representaive vowel, glide, and diphthong segments ....................... 37
3.6 Percentages of vowel, glide, and diphthong segments with maximum $L_{X_k}^{Rayl}$, $L_{X_k}^{Exp}$, and $L_{X_k}^{\gamma}$ values. .......................................................... 38
3.7 Average $v_k$ (shape parameter of the Gamma PDF) over the representative nasal segments in the speech sample. ......................... 39
3.8 Percentages of nasal segments with maximum $L_{X_k}^{Rayl}$, $L_{X_k}^{Exp}$, and $L_{X_k}^{\gamma}$ values .......................................................... 40
3.9 Average $v_k$ (shape parameter of the Gamma PDF) over the representative fricative and affricate segments in the speech sample. .......... 41
3.10 Percentages of fricative and affricate segments maximum $L_{X_k}^{Rayl}$, $L_{X_k}^{Exp}$, and $L_{X_k}^{\gamma}$ values .................................................. 42
4.1 Gain function of the Rayleigh-MMSE estimator as a function of $\xi$. 57
4.2 Gain function of Rayleigh-MMSE estimator as a function of $\gamma$. 57
4.3 Gain function of the Gamma-MAP estimator as a function of the a-posteriori SNR. .......................................................... 61
4.4 Gain function of the Gamma-MAP estimator as a function of the a-priori SNR. .......................................................... 62
4.5 Gain function of the Gamma-MMSE estimator as a function of the a-posteriori SNR. .......................................................... 65
4.6 Gain function of Gamma-MMSE estimator as a function of the a-priori SNR. .......................................................... 66
4.7 The mean SIG, BKG, and OVL scores obtained from the listening test. 71
4.8 Waveforms of 0-dB white-noise corrupted speech processed by Wiener, Rayleigh-MMSE, Gamma-MAP, and Gamma-MMSE estimators ... 72
4.9 Signal distortion values and noise reduction factors of the enhanced speech samples as a function of SNR. 74

5.1 The highpass filter used to shape a fricative frame with a highpass spectrum trend. 79

5.2 Block diagram of the proposed class-based noise reduction system. 82
Chapter 1

Introduction

1.1 Background and Motivation

Speech can be described as the verbal form of human communication. Technological advances, such as the development of telephone and radio, enhance our speech communication capability even more by enabling direct communication across larger distance, broader audiences, and more challenging circumstances. Speech signal processing aims to enhance speech communication systems by manipulating speech signals at a point in the communication link\(^1\) so that the conveyed message can be recovered with minimum loss on the receiver’s side. Some examples of speech signal processing are speech coding, background noise reduction, time-scale or pitch modification, and speech recognition.

Noise reduction systems aim to enhance the quality of noisy speech by suppressing the background noise. However, there is always a tradeoff between noise reduction

\(^1\)Speech communication link refers not only to human-to-human communication but also human-to-machine or machine-to-human communication
and signal distortion, i.e., the speech signal itself becomes more distorted when more background noise is removed. A study that compared the intelligibility of some noise-reduced speech with that of the noisy speech can be found in [23], and it showed that most noise reduction algorithms rarely improve, or even lower, the intelligibility\(^2\) of noisy speech. This is because the acoustic cues needed to understand the message are either distorted or masked by the residual noise. We believe that in designing speech processing algorithms, the changes to acoustic cues induced by the processing should be taken into account.

The field of acoustic-phonetics studies the acoustic characteristics of speech sounds. It observes the measurable acoustic properties that occur in certain linguistic units (e.g., phonemes, syllables, words) and studies how their changes affect speech perception. We found that speech enhancement systems, with the exception of speech recognition systems, do not generally incorporate acoustic-phonetic knowledge in their developments.

In this thesis, we aim to design a single microphone noise reduction system that utilizes information from acoustic-phonetics to enhance the intelligibility of noisy speech. In our view, this objective requires us to consider the following aspects:

1. What are the cues that should be enhanced?
2. How should we incorporate this information in the design of the noise reduction system?
3. What is the noise reduction algorithm that we should use?

To address the first question, we performed a literature study on the acoustic characteristics of speech signals. The acoustic properties of speech sounds in each

\(^2\)the percentage of correct identification of the intended message
phonetic class, as well as the discriminating cues of sounds within a class, are summarized in Chapter 2. At the end of Chapter 2, we use this information to design a class-based noise reduction algorithm coupled with some acoustic cue enhancement strategies.

We additionally assumed that the system has the sound class knowledge of the input speech. In practice, only the noisy speech and the noise power spectrum estimate are available, so that the classification must be obtained from these pieces of information. Developing a sound classification system that performs as well as human speech recognition performance is quite a challenging task. However, when the SNR level is not too low, the reliability of an automated classification system might be sufficient to benefit those with hearing impairments. Methods to classify noisy speech will not be explored in this thesis. We assume that the class information is known, and we obtain it by performing manual labelling and segmentation.

We employ a statistical-model based (Bayesian) Short-Time Fourier Transform (STFT) domain algorithm, similar to the one proposed by Ephraim and Malah in [16], for the core noise suppression algorithm. This algorithm modifies only the spectral amplitude of the noisy speech and uses the noisy phase as the phase estimate. It also assumes that the joint probability distribution function (PDF) of speech and noise is explicitly known. In Chapter 3, we evaluate the model assumptions for this algorithm. We evaluate whether the speech Short-Time Spectral Amplitudes (STSAs) in different speech classes are better modelled by different parametric PDFs. In Chapter 4, the derivation of the STFT estimators based on the conclusion of Chapter 3 is presented, and the subjective and objective performances of the estimators are evaluated. Based on this evaluation, we propose to apply different estimators for different classes of
speech. The system implementation and the simulation results of the proposed class-
based noise reduction system are then summarized in Chapter 5.

The next three sections summarize the objectives, contributions, and outline of
this thesis, respectively.

1.2 Thesis Objectives

The main objective of this thesis is to design a single microphone noise reduction
system that aims to recover speech intelligibility by utilizing both acoustic-phonetic
and statistical knowledge. To achieve this objective, the following tasks will be un-
dertaken:

1. Perform a literature study on the acoustic-phonetics properties of speech sounds,
and summarize the acoustic cues that are known to play roles in speech recog-
nition.

2. Evaluate the parametric PDF models of speech STSA in different sound classes.
Observe if there are any patterns that characterize the PDFs of speech STSAs
in different sound classes.

3. Develop speech STSA estimators which use the PDF models concluded from
the previous study, and evaluate the subjective and objective performances of
these algorithms.

4. Propose a noise reduction algorithm, based on the results of acoustic-phonetic
and statistical-model studies, and evaluate its intelligibility performance.
1.3 Thesis Contributions

The contributions of this thesis are:

• A design proposal for a single microphone noise reduction system which uses both acoustic-phonetics and statistical knowledge to improve the intelligibility of noisy speech.

• A study of the PDF models of the STSA of unclassified and classified speech. The signal model is similar to that proposed by Ephraim and Malah in [16]. The STSAs are assumed to be independent in time and frequency. Furthermore, the STSAs at all frequency-bins and frame indices are assumed to have an identical parametric PDF with varying parameters.

• The derivation of Bayesian STSA estimators (the Gamma-MAP and Gamma-MMSE estimators), which employ the Gamma PDF assumption for the speech STSA, and the evaluation of their performances.

1.4 Thesis Outline

The rest of the thesis is organized as follows:

Chapter 2 provides a short background on speech perception and a summary of literature review of the characteristics of speech sounds. A summary of the cues that characterize each sound class as well as the effects of manipulating them are presented. Following the literature review, we present a proposal for a class-based processing scheme, which couples a statistical noise reduction method with a class-based processing strategy that directly modifies the acoustic cues known to be important.
for speech intelligibility.

**Chapter 3** presents the procedures and results of the evaluations of the PDF models for unclassified and classified speech STSA.

**Chapter 4** presents a literature review of noise reduction systems. In particular, we present in more detail a type of Bayesian STSA estimator which operates in the STFT domain. Based on the results of the PDF study in Chapter 3, two STSA estimators (Gamma-MAP and Gamma-MMSE), which assume a Gamma distribution for the speech STSA, are derived and their subjective and objective performances are evaluated.

**Chapter 5** presents the implementation details of the proposed class-based noise reduction system, as well as the intelligibility evaluation of the output speech.

**Chapter 6** provides a summary and the conclusions of this thesis, as well as some suggestions for future work.
Chapter 2

Background and Proposed System

Overview

This chapter provides some background on speech perception and acoustic-phonetics, as well as an overview of the proposed noise reduction system. Section 2.1 presents a short introduction to speech perception. Section 2.2 provides some background on the characteristics of speech sounds, acoustic-phonetic classification, the acoustic cues that characterize each sound class, and the effects of cue manipulation on phoneme recognition. Section 2.3 outlines the proposed class-based noise reduction system. The chapter summary is finally presented in Section 2.4.

2.1 Overview of Speech Perception

A speech signal is transmitted in the form of acoustic waves. Our sensory receptors pick up the physical stimuli and code this information in the form of nerve impulses. Our perception process translates these neural stimuli into meaningful linguistic units.
2.1. OVERVIEW OF SPEECH PERCEPTION

(e.g., phonemes, morphemes, words, and sentences).

To study speech, researchers break down speech into hierarchical linguistic units. The lowest element in the hierarchy is the phoneme [39, Ch.9], followed by the morpheme, the smallest unit that has meaning in language. Morphemes are then combined to construct words, and words are combined to construct sentences. Syntax (grammar) provides the rules of how these units can be combined, and semantics provide the context of how these units are combined to create meaning. While some researchers do not even agree that a phoneme can be seen as a perceptual unit, speech analysis in terms of phonemes has been widely used [33, p.247].

Speech perception is not a hierarchical process—it is both a bottom-up and top-down procedure. However, the experiment by Fletcher, as cited in [2], shows that in the absence of a context, speech recognition starts by the identification of phones\(^1\), then syllables, and then words, and so on. Fletcher conducted the articulation\(^2\) experiment using a vocabulary set of nonsense CVC, CV, and VC syllables (C: consonant, V: vowel), and found that for CVC syllable articulation, the following relation holds

\[
S(\alpha) \approx c^2(\alpha)v(\alpha),
\]

where \(S\) is the probability of correctly identifying the nonsense CVC syllable, \(c\) and \(v\) are the probabilities of correctly identifying the consonant and vowel phones, and \(\alpha\) denotes the signal-to-noise ratio. Furthermore, Fletcher and Steinberg (as cited in [2]) were able to find a relationship between nonsense CVC syllables and word recognition rate in terms of the entropy of the word corpus, which was empirically determined.

\(^1\)an instance of a phoneme

\(^2\)probability of correct recognition when context is not present
2.1. OVERVIEW OF SPEECH PERCEPTION

How are different acoustic patterns mapped into phonemes? The mapping is not so simple that a set of acoustic patterns can be mapped to a phoneme in a one-to-one manner. Different acoustic patterns may correspond to the same phoneme. This is because phoneme production is influenced by many factors, such as adjacent phonemes, speaking rate, and variability among speakers. However, a model of speech perception proposes that it should be possible to find relatively invariant acoustic cues that correspond to a certain perceived phoneme (see [33, ch.8, Sec. 4B and 5]). Cole and Scott [11] noted that speech perception involves simultaneous identification of different types of acoustic cues: invariant cues, context-dependent cues (acoustic cues that vary, for example, due to coarticulation), and cues provided by the time waveform, such as periodicity or abruptness of amplitude changes. Speech enhancement algorithms should take into account how signal processing affects these cues and how they affect speech perception.

Speech enhancement techniques aimed at improving intelligibility target the key acoustic cues which are known to be crucial for speech recognition. An early method that proposes to improve the intelligibility of noisy speech can be found in [12]. The proposed system is not intended to enhance the SNR, and noise reduction is not attempted. It works as follows: If the initial sound of the syllable is /s/, then the syllable is passed through a highpass filter (1-10 kHz). Syllables that start with a stop consonant are altered by inserting a short pause before the sound appears. The experimental results show a significant intelligibility improvement, especially in low SNR conditions (about 25%, 20%, and 5% improvement in -8, -4, and 0 dB SNR, respectively). Some methods that enhance the distinctive acoustic features in clean speech can be found in [21, 22]. The methods alter the signal, so that interclass
and intraclass acoustic cues are amplified or contrasted. Significant enhancement in intelligibility was also reported.

Noise reduction algorithms enhance the SNR of the signal; therefore, they are expected to improve speech intelligibility. However, there is always a tradeoff between noise reduction and signal distortion—the more noise suppression is applied, the more signal distortion is incurred. As reported in [24], most noise reduction algorithms proposed in the literature fail to improve, or even reduce, the intelligibility of noisy speech.

In this work, we aim to compare several noise reduction methods in terms of signal distortion and noise reduction. We propose to apply a noise reduction algorithm that causes minimum signal distortion around the cue-rich regions in the signal, and apply a more aggressive suppression otherwise. Moreover, we also apply class-based cue-enhancement processing, similar to some of those proposed in [21]. In the next section, a background on the acoustic characteristics of speech sounds, as well as a summary of the acoustic cues in each sound class is presented. Section 2.3 then presents a general overview of the class-based processing scheme.

2.2 Acoustic Characteristics of Speech Sounds

The major organs that are responsible for speech production are the lungs, larynx, and the vocal tract. The lungs provide a source of airstream that is pushed upwards into the larynx. This airstream is modulated by the vocal folds, which are located in the larynx, as well as the articulators located in the oral cavity. The modulated airflow is filtered according to the shape of the vocal tract. The vocal tract consists of the nasal and oral cavities, and the articulators inside the oral cavity (lips, tongue,
2.2. ACOUSTIC CHARACTERISTICS OF SPEECH SOUNDS

teeth, velum, palate, etc.).

Depending on the state of the vocal folds, speech sounds can be categorized into voiced and unvoiced [41, Ch.3]. In the voiced state, the vocal folds oscillate between opening and closing states. As a result, the airstream from the lungs is modulated in a periodic-like manner, and the spectra of these sounds are characterized by harmonics. The fundamental frequency of this oscillation (also known as pitch frequency) typically ranges between 75 to 300 Hz. Males usually have more massive vocal folds; therefore, they tend to have low pitch frequencies. In the unvoiced state, the vocal folds are more tense and closer together, and turbulence occurs at the vocal cords. The spectrum can be characterized as noise-like.

Vowels are characterized by a longer duration in which the signal is stationary, while consonants are characterized by a rapid change in the spectrum over a time period of 10-30 ms [33, p.296]. Glides and diphthongs are moving vowels. They are characterized by a rapid transition from one vowel to another. Consonants can be further classified into plosives, fricatives, liquids, affricates, and nasals. More detailed discussions on the acoustic characteristics of each sound class are presented in the following subsections. Phonemes in each sound category can be seen in Table 2.1.

2.2.1 Vowels

In the production of vowel sounds, the vocal folds vibrate, and the vocal tract configuration is more steady compared to that during the production of consonant sounds. Different vowels correspond to different configurations of the vocal tract. The acoustic waveform is characterized by a high energy level as well as periodicity. The spectrum
<table>
<thead>
<tr>
<th>Vowels</th>
<th>Consonants</th>
</tr>
</thead>
<tbody>
<tr>
<td>/i/</td>
<td>see, bean</td>
</tr>
<tr>
<td>/e/</td>
<td>pin, mill</td>
</tr>
<tr>
<td>/a/</td>
<td>about, turn</td>
</tr>
<tr>
<td>/æ/</td>
<td>pat, sand</td>
</tr>
<tr>
<td>/ɛ/</td>
<td>met, den</td>
</tr>
<tr>
<td>/ʌ/</td>
<td>sung, done</td>
</tr>
<tr>
<td>/ʊ/</td>
<td>pin, mill</td>
</tr>
<tr>
<td>/o/</td>
<td>about, turn</td>
</tr>
<tr>
<td>/ʌ/</td>
<td>pat, sand</td>
</tr>
<tr>
<td>/ɛ/</td>
<td>met, den</td>
</tr>
<tr>
<td>/ʌ/</td>
<td>sung, done</td>
</tr>
</tbody>
</table>

Voiceless vowel: /h/ he

Diphthongs:
| /ej/      | pay               |
| /aj/      | lie               |
| /ɔj/      | boy               |
| /æj/      | pay               |
| /ow/      | row               |
| /aw/      | cow               |

Plosives:
- Voiced:
  - /b/ bun
  - /d/ dole
  - /g/ gale
- Unvoiced:
  - /p/ pit
  - /t/ tan
  - /k/ can

Approximants:

Liquids:
| /l/      | land |
| /w/      | we   |

Glides:
| /j/      | you  |

Fricatives:
- Voiced:
  - /v/ van
  - /θ/ then
  - /z/ zoo
  - /ʒ/ leisure
- Unvoiced:
  - /f/ fan
  - /θ/ thin
  - /s/ sue
  - /ʃ/ ship

Nasals:
| /m/      | man  |
| /n/      | can  |
| /ŋ/      | sing |

Affricates:
| /tʃ/     | church |
| /dʒ/     | judge  |

Table 2.1: Phonemes in American English [42, Ch.2]
2.2. **ACOUSTIC CHARACTERISTICS OF SPEECH SOUNDS**

is characterized by the presence of harmonic frequencies and formants, as shown in an example in Figure 2.1. Formants are peaks in the spectrum, as shown by $F_1 - F_3$ in the spectrum plot. These are the resonant frequencies of the vocal tract. Vowels are perceptually discriminated by the relative locations of the first three formant frequencies [42, p.153].

![Figure 2.1: Waveform, spectrogram, and spectrum of vowel /i/ (eve).](image)

**2.2.2 Glides, Diphthongs, and Liquids**

These classes of sounds are vowel-like in nature and characterized by transitional sounds. Glides can be viewed as moving vowels; they move from one vocal tract configuration to another. A diphthong is similar to a glide, and it can be viewed
as a sequence of a vowel and a glide. The transition rate is faster in glides than in diphthongs. Liquids are similar to glides but they are produced by the tongue forming side branches in the oral tract, which introduce anti-resonances [41, p.93-94]. Formant transitions in liquids, glides, and diphthongs are smooth and continuous, except in liquid /l/, since the tongue tip touches the palate and releases before the next vowel [41, p.94].

2.2.3 Plosives

The waveform of plosive sounds can be characterized by the following sequence: (1) a complete stop (unvoiced stop) or voice bar (voiced stop), (2) a burst followed by an aspiration, and (3) the onset of the next vowel. This sequence is illustrated in Figure 2.2. Voice bar occurs when the phoneme is voiced, but it does not necessarily appear. Furthermore, the duration between the release of the burst and the start of the voicing of the next vowel (the voicing onset time (VOT)) is shorter in voiced stops compared to that in unvoiced stops. A contrasting illustration between voiced and unvoiced plosives can also be seen in Figure 2.2.

Some of the acoustic characteristics that influence the perception of plosive consonants are:

1. Abruptness of amplitude change at the onset of the burst (sudden fall in amplitude at the closure). The amplitude rise occurs at all frequencies. This property is also present in the characteristics of nasal consonants, but the amplitude contrast is higher in voiced plosives compared to that in nasal consonants. Manipulation of the relative amplitude of the nasal murmur and the following vowel shifts listeners’ perception from nasal consonant to plosive, when the relative
2.2. ACOUSTIC CHARACTERISTICS OF SPEECH SOUNDS

Figure 2.2: Waveforms of unvoiced and voiced plosives: (a) /k/ in ’coal’ and (b) /g/ in ’goal’.

amplitude is modified to -8 to -15 dB [43, p.28]. A sudden amplitude increase can also be found in a fricative followed by a vowel, but the amplitude increase does not occur in all frequencies. Moreover, the rate of amplitude change is more abrupt in plosives compared to that in fricatives [43, p.26].

2. Presence or absence of voicing, shortly (up to about 20 ms) before and after the burst release. In unvoiced plosives, voicing does not start within 20 ms following the burst release, and the consonant is aspirated. In an experiment described in [43, p.31], listeners hear an unvoiced plosive if low frequency periodicity does not appear within 20 ms of the burst release, but hear a voiced plosive otherwise. Furthermore, the relative amplitude of the aspiration and the next
vowel influences the voiced-unvoiced perception. Louder aspiration tends to shift the perception to unvoiced (also in [43, p.31]).

3. The spectrum trend at the period between the burst release and the onset of the next vowel, which is determined by the place of articulation [43, p.6-22], i.e., the highest contact point of the articulators in the vocal tract during the constriction. The three common places of articulation are at the labial (lips), alveolar (the ridge just behind the upper teeth), and velar (the soft palate at the rear portion of the roof of the mouth) surfaces. Labial plosives are /p/, /b/. Alveolar plosives are /t/, /d/. Velar plosives are /k/, /g/. The theoretical spectrum trends according to the place of articulation are [43, p.10]:

- Labial: the spectrum is rising with increasing frequency.
- Alveolar: the spectrum is flat or falling with increasing frequency.
- Velar: the spectrum is compact or characterized by a prominent spectral peak in the mid-frequency range.

### 2.2.4 Fricatives

Fricatives are produced by creating a narrow constriction in the oral tract and passing air through the constriction. They are characterized by a highpass noise-like spectrum, but have lower energy level compared to that of vowels. In unvoiced fricatives, the vocal folds do not vibrate, while in voiced fricatives, the vocal folds vibrate simultaneously with the noise generation at the oral tract. Figure 2.3 displays the waveforms of unvoiced and voiced fricatives /s/ (sue) and /z/ (zoo).

Fricatives are identified by the following:
2.2. ACOUSTIC CHARACTERISTICS OF SPEECH SOUNDS

Figure 2.3: Waveform, spectrogram, and spectrum plots of unvoiced fricative /s/ in ’sue’ (left column) and voiced fricative /z/ in ’zoo’ (right column). The fricatives, /s/ and /z/, start and end at approximately 0.02 and 0.2 s. Spectrum plots in row 3 are taken using 42-ms Hanning window centered at 0.1 s.

1. The gross spectrum trend, which is determined by the place of articulation [41, p.88]. /s/, which has a constriction point at the palate, has a highpass spectrum, while /f/, which has a constriction point at the upper-teeth and lips, has a flat or mild upward trend.

2. The duration of the frication prior to the onset of voicing of the next vowel. The duration is shorter for voiced fricatives and longer for unvoiced fricatives [41, p.88].
Furthermore, the intelligibility enhancement experiments in [12] and [21] showed that fricative intelligibility is significantly improved by amplitude amplification of the frication segment.

### 2.2.5 Nasals

Nasals are produced by completely closing the oral tract, letting an airstream flow through the nasal cavity. The vocal folds are in the voicing state. The waveform is characterized by periodicity and an abrupt amplitude decrease during the oral tract closure. The opening of the nasal passage produces extra resonances and the closing of the oral tract produces an antiresonance [41, p.84]. The resulting effect produces a spectrum that is characterized by high energy in the low-frequency range and weak energy in the high-frequency range, as shown in Figure 2.4.

Some of the acoustic properties that characterize nasal sounds are:

1. The duration of the nasal murmur and the relative amplitude of the nasal murmur, compared to the amplitude of the following vowel. By reducing the duration and decreasing the relative amplitude, listeners’ perceptions shift from nasal sounds to voiced stop consonants [43, p.27-28].

2. The pattern of formant ($F_2$) transition that precedes and follows the closure of the oral constriction [41, p.84], [42, p.163].

### 2.2.6 Affricates and Whisper

We will not discuss the acoustic characteristics of affricates and whisper in detail. Affricates can be viewed as a sequence of stop, burst, and fricative. /h/ is the only
2.2. ACOUSTIC CHARACTERISTICS OF SPEECH SOUNDS

Figure 2.4: Waveform, spectrogram, and spectrum plots of nasal sounds /m/ in ‘me’ (left column) and /n/ in ‘need’ (right column). /m/ and /n/ occur from approximately 0.05 to 0.2 s (/m/) and 0.05 to 0.15 s (/n/). Spectrum plots in row 3 are taken using 42-ms Hanning window centered at 0.12 s (a) and 0.1 s (b).

whisper in the English language. It is produced by creating a turbulence at the glottis. The glottis is open, and the vocal folds are not vibrating. The configuration of the oral tract is the same as the following vowel. In the proposed noise reduction system, affricate is treated as a combination of a burst followed by a fricative segment, and a whisper is treated similarly to a vowel.
2.3. Overview of the Proposed System

To perform the noise reduction, we employ a Bayesian noise reduction algorithm which operates in the STFT domain. The algorithm is based on a statistical model which assumes that the a-priori distribution of the clean speech STFT magnitude is known. The block diagram of this algorithm is shown in Figure 2.5.

The processing is performed in a frame-based manner, where the frame duration and shift rate are identical for all frames. $y(n)$ and $x(n)$ denote the noisy and clean speech, respectively. $\vec{y}(m) = [y((m-1)N + 1) \cdots y(mN)]$, and similarly, $\vec{x}(m)$, denote the $m$-th noisy and clean frames, respectively, where $N$ denotes the frame length in number of samples. $Y_k(m)$ and $e^{j\theta_k(m)}$ denote the magnitude and phase of the $k$-th noisy STFT component, and $\hat{X}_k(m)$ denotes the STFT magnitude of the estimated clean signal. $k = 1, \cdots, N$ denotes the frequency-bin index, and the argument $m$ inside the parentheses denotes the frame index.

The spectral amplitude estimation block is explored in Chapters 3, 4, and part of Chapter 5. In Chapter 3, we explore the assumptions on the PDF of speech STFT magnitude. In Chapter 4, two estimators that are based on the conclusion obtained in Chapter 3 are derived, and their performances, in terms of signal distortion and noise reduction, are evaluated. Based on this evaluation, we propose to use different
estimators for different speech classes. In Chapter 5, we comment on the simulation results of the proposed class-based estimators.

In addition to the spectral amplitude estimator, we also propose to apply some class-based cue-enhancement processing, similar to those performed in [21]. The signal modification targets directly some of the invariant cues that are known to play roles in speech recognition. We assume that the system has the class information of the noisy input. In our work, we use speech samples that have been manually classified, since we only aim to evaluate the proposed class-based strategy in terms of its intelligibility performance. The class-based cue enhancement strategy is described as follows.

For vowels, glides, diphthongs, and whispers, no additional processing is performed other than the spectrum amplitude estimation. For plosives, as well as affricates, we assume that the stop, burst, and aspiration portions of the signals are known. Strong suppression is performed to the frames tagged as stop. Prior to the noise reduction block, burst and aspiration frames are analyzed to obtain their gross spectrum trends, and based on this information, spectral shaping is applied to the enhanced spectrum. Furthermore, burst segments are amplified by +6 dB, and aspirated segments are amplified by +3 dB. Similarly for fricatives, the noisy spectrum is analyzed to determine whether it has a highpass or flat spectrum, and spectral shaping is applied to the enhanced spectrum. The processed frame is then amplified by +3 dB. For nasals, the enhanced spectrum amplitude is amplified by +3 dB. An affricate is processed as a sequence of burst and fricative segments. The procedure is summarized in Figure 2.6.

The detailed implementation of the proposed system is discussed in Chapter 5.
2.4 Summary

In this chapter, we have presented a short background on the acoustic characteristics of speech sounds, as well as a proposal for a class-based noise reduction system. Section 2.1 provides a short introduction on speech perception. The important points are summarized as follows:

- Speech perception is not a hierarchical procedure, i.e., to recognize speech, we utilize cues in different contextual levels, e.g., acoustic cues, syntactic cues, semantic cues, etc., and there are interactions between different recognition layers. However, the articulation study performed by Fletcher, as cited in [2], showed that when contexts are removed, speech recognition begins by decoding independent cues available in the acoustic signal.
2.4. SUMMARY

• A model of speech perception suggests that it is possible to find relatively invariant cues that signal a specific phoneme [33, p.294]. Several proposed intelligibility enhancement algorithms that enhance these cues, e.g., in [12, 21, 22], have shown significant intelligibility improvement.

Section 2.2 discusses the acoustic characteristics of each sound class and some invariant cues that signal the phonemes in each sound class. Section 2.3 outlines the proposed class-based noise reduction algorithm (the implementation details are discussed in Chapter 5). The core noise suppression algorithm is statistical-based, and it works in the STFT domain (Figure 2.5). We also proposed to apply a class-based cue enhancement strategy, similar to some of those performed in [21]. This proposed class-based processing is summarized in Figure 2.6.
Chapter 3

The PDF of Unclassified and Classified Speech STSA

This chapter discusses the PDF models of speech STSA used in the core noise reduction algorithm. We evaluated the goodness-of-fits of several parametric PDFs to the empirical speech data. The PDF models of both unclassified and classified speech STSA are evaluated. Section 3.1 provides an overview of the model; model assumptions and the procedures used to evaluate the PDF fit are described. Section 3.2 discusses the model evaluation of unclassified speech, and Section 3.3 discusses the model evaluation of classified speech. Section 3.4 provides the summary and conclusions of this chapter.

3.1 Overview

Since speech has a nonstationary nature, its processing is usually performed in a frame-based manner. That is, the input data are processed in chunks. Each chunk
3.1. OVER VIEW

usually corresponds to 16 to 35 ms of speech, during which the signal is approximately stationary. In our analysis, we chose a fixed 32-ms frame duration and a 75-percent frame overlap. In addition, the processing was performed in the spectral domain—the discrete Short-Time Fourier Transform (STFT) was applied to each frame. Furthermore, it is also a common practice in spectral domain speech enhancement methods to modify only the magnitude of the STFT coefficients (the short-time spectral amplitude (STSA)). The noisy STFT phase is left unmodified and combined with the modified STSA to produce the enhanced speech.

Let us denote the time-domain speech samples by $x(n)$, where $n$ denotes the time index. The STFT coefficients are denoted by the following notations:

$$
\mathcal{F} \{x((m-1)N + 1), x((m-1)N + 2), \ldots, x(mN)\} \rightarrow \{X_1(m)e^{j\theta_{X,1}(m)}, X_2(m)e^{j\theta_{X,2}(m)}, \ldots, X_N(m)e^{j\theta_{X,N}(m)}\},
$$

(3.1)

where $\mathcal{F}$ denotes the Discrete Fourier Transform (DFT) operation, $m$ the frame index, $N$ the number of samples in a frame, $X_k(m)$ the STSA in frequency-bin $k$ and frame index $m$, and $\theta_{X,k}(m)$ the STFT phase in frequency-bin $k$ and frame index $m$.

The core noise reduction algorithm employs a statistical model similar to that proposed in [16]. It models the speech and noise STFT coefficients in each frequency-bin $k$ and frame-index $m$ with complex zero-mean circularly symmetric Gaussian random variables. Speech and noise are assumed to be independent, and their STFT coefficients are also assumed to be independent in both frequency and time. Noise reduction algorithms which are based on a similar model, but employ different parametric PDF assumptions have been proposed in different papers (Chapter 4, Table 4.2). Due to the complex circularly symmetric Gaussian assumption, the STFT amplitude and phase are independent and have Rayleigh and Uniform distributions,
3.1. **OVERVIEW**

respectively. However, this is not true for non-Gaussian complex random variables, which complicates the derivation of the estimator. We proposed an algorithm which assumes a similar signal model as in [16], with an additional assumption that the speech STFT amplitude and phase are independent. This chapter evaluates the PDF assumption of the speech STSA. The noise reduction algorithms derived based on the results of this chapter are described in Chapter 4.

In [16], the signal parameter is estimated by smoothing the parameter estimates obtained from the current as well as past estimates (4.18). This indicates an assumption that $X_k$ is approximately stationary for a number of frames. We are interested in knowing whether $X_k(m)$ was better modeled by a Rayleigh PDF, as assumed in [16], or by other parametric distributions. Moreover, the PDF models for both unclassified and classified STSAs are evaluated, where the classes correspond to the sound categories discussed in Chapter 2. The following PDFs were considered:

- **Rayleigh PDF with parameter $a$:**

  $$f^{\text{Rayl}}(x; a) = \frac{x}{a} e^{-\frac{x^2}{2a}}, a > 0, x \geq 0. \quad (3.2)$$

  Given $N_f$ realizations of the r.v. $X(i), i = 1, 2, \cdots, N_f$, the Maximum Likelihood Estimate (MLE) of $a$ is one that maximizes the following log-likelihood criterion:

  $$\hat{a} = \max_a \sum_{i=1}^{N_f} \log f^{\text{Rayl}}(X(i); a), \quad (3.3)$$

  Substituting (3.2) into (3.3) and solving it gives

  $$\hat{a} = \frac{1}{2N_f} \sum_{i=1}^{N_f} X^2(i). \quad (3.4)$$

- **Exponential PDF with parameter $\lambda$:**

  $$f^{\text{Exp}}(x; \lambda) = \lambda e^{-\lambda x}, \lambda > 0, x \geq 0. \quad (3.5)$$
3.1. OVERVIEW

The corresponding MLE of $\lambda$ is

$$\hat{\lambda} = \left( \frac{1}{N_f} \sum_{i=1}^{N_f} X(i) \right)^{-1}. \quad (3.6)$$

- Gamma PDF ($\gamma$-PDF) with shape and scale parameters $v$ and $\zeta$, respectively:

$$f^\gamma(x; v, \zeta) = \frac{x^{v-1}}{\Gamma(v)\zeta^v} e^{-\frac{x}{\zeta}}, \quad v > 0, \zeta > 0, \quad x \geq 0, \quad (3.7)$$

Note that the Exponential PDF belongs to the $\gamma$-PDF family with $v = 1$. The MLE of $\zeta$, assuming $v$ is known is straightforward:

$$\hat{\zeta} = \frac{1}{vN_f} \sum_{i=1}^{N_f} X(m). \quad (3.8)$$

The estimation of $v$, however, is not so straightforward and there is no known closed-form MLE solution. $\hat{v}$ is solved by substituting (3.8) into the log-likelihood function (as in (3.3)), and using MATLAB to numerically solve for $v$ that maximizes this function.

The Rayleigh, Exponential, and Gamma PDFs are depicted in Figure 3.1. Note that the Exponential PDF has a heavier tail compared to that of the Rayleigh PDF. Furthermore, given equal second moments, varying $v$ is equal to controlling the lobe width or tail of the Gamma PDF—higher $v$ means narrower lobe width and lighter tail. When $v \to \infty$, the limit of the function goes to the Dirac delta function.

To evaluate how well each PDF candidate fits the empirical data, we use the following measures:

- Log-Likelihood value:

$$L = \sum_{i=1}^{N_f} \log f(X(i)). \quad (3.9)$$

A large $L$-value indicates that the model represents the data well.
• Chi-squared value:

\[
\chi^2 = \sum_{i=1}^{K} \frac{n_i - (N_f \times p_i)}{(N_f \times p_i)},
\]

(3.10)

where \( K \) is the number of bins, \( n_i \) is the number of observations that fall into bin-\( i \), and \( p_i \) is the probability that a sample falls into bin-\( i \). The \( \chi^2 \)-value is the test statistic used in the Pearson’s \( \chi^2 \) goodness-of-fit test, which evaluates the hypothesis of whether observation data come from a specified distribution. In general, a smaller \( \chi^2 \)-value indicates a better fit.

The Log-Likelihood functions of Rayleigh, Exponential, and Gamma PDFs are denoted as \( \mathcal{L}^{\text{Rayl}} \), \( \mathcal{L}^{\text{Exp}} \), and \( \mathcal{L}^{\gamma} \), respectively. Similar notations are also employed to denote the \( \chi^2 \) functions of Rayleigh, Exponential, and Gamma PDFs.

In the following two sections, the PDF models for the unclassified (Section 3.2) and classified (Section 3.3) speech STSAs are discussed. Based on the proposed measures, we show that different models should be used for unclassified and classified speech. Section 3.4 provides summary and conclusions to the chapter.
3.2 The PDF of Unclassified Speech Spectral Amplitude

The speech sample consists of approximately 1.5 minutes of continuous speech, spoken by a female native English speaker. Recording was performed in a quiet room, using a desktop microphone placed approximately 10 cm in front of the speaker’s lips. The original sampling rate was 32 kHz, and the file was resampled to 16 kHz. The time-domain sample was transformed into the STFT domain using a 512-point (32-ms) normalized Hanning window and 75% frame overlap. The normalized Hanning window has the following expression:

\[ w(k) = \sqrt{\frac{1}{6}} \left( 1 - \cos \left( \frac{2\pi k}{N} + \frac{\pi}{N} \right) \right), \quad (3.11) \]

where \( k = 0, \cdots, N - 1 \), and \( N \) denotes the window length (512). It can be verified that \( \sum_{m=0}^{\infty} w^2(m \frac{N}{4} + k) = 1 \), i.e., the sum of overlapping windows at each time instant \( k \) is unity.

The evaluation of parametric PDF models for all \( X_k(m), k = 0, \cdots, 256 \), is conducted as follows. The STFT-domain signal is organized into segments, where each segment consists of \( N_f \) frames. Consecutive segments are also spaced by \( N_f \) frames (no segment overlap). Furthermore, silence intervals were discarded, and all \( N_f \) frames within a segment are immediate neighbours (uninterrupted by silence). We assume that a segment is stationary, i.e., \( X_k(m;s) \), where \( m = 1, \cdots, N_f \) and \( s \) denotes segment index, are independent and identically distributed (iid). For each segment \( s \), the Rayleigh, Exponential, and Gamma MLE parameters for all frequency bins are calculated. These MLE parameters are then substituted into (3.2)-(3.7), and subsequently used to calculate \( \mathcal{L}^{\text{Rayl}}_{X_k}(s) \), \( \mathcal{L}^{\text{Exp}}_{X_k}(s) \), and \( \mathcal{L}^{\gamma}_{X_k}(s) \), the Log-Likelihood values of
Rayleigh, Exponential, and Gamma PDFs for each $X_k$. The $\chi^2$-value of each candidate PDF is also calculated in the same manner. For each segment and frequency bin, the Log-Likelihood and $\chi^2$ values of the candidate PDFs are compared. The proportions of segments which have maximum $L_{X_k}^{\text{Rayl}}(s)$, $L_{X_k}^{\text{Exp}}(s)$, or $L_{X_k}^{\gamma}(s)$ (minimum $\chi^2$ values) are plotted, for example, as in Figure 3.3.

We observed that ML parameters of the Gamma PDF often results in log-likelihood values that are greater than those of one-parameter PDFs. This is because the Gamma PDF has two degrees of freedom, whereas the other models merely have one degree of freedom. Therefore, we fixed $v_k$, one of the two parameters of Gamma PDF, to a constant value. To choose $v_k$, we first evaluated the MLE of $(v_k, \zeta_k)$ for all segments in the sample, and averaged $v_k$ over all the segments. For $N_f = 5, 10, 20, 30$, the average $v_k$ is plotted in Figure 3.2. Based on these plots, we fixed $v$ to be 8, 3, 2, 1 for $N_f = 5, 10, 20, 30$, respectively. Note that $v$ decreases as $N_f$ increases. $N_f$ is related to the smoothing factor used in the decision-directed parameter estimator proposed in [16] (4.18). Increasing $N_f$ indicates that more frames are used to estimate the current frame’s parameters, which is similar to the effect caused by increasing the smoothing factor. The plot only shows the 800-8000 Hz frequency range since $v_k$ takes large values in 0-800 Hz range, especially for $N_f = 5$ (in the order of $10^2$).

We evaluated the Log-Likelihood values for $N_f = 5, 10, 20$, and 30. Since there are too few samples, $\chi^2$ values for $N_f = 5$ and 10 were not evaluated. The percentages of segments with maximum Rayleigh, Exponential, and Gamma Log-Likelihood values are plotted in Figure 3.3, for $N_f = 5,10,20$, and 30, respectively. As shown in the figures, the Rayleigh PDF is preferred for $N_f = 5$, but the Exponential PDF becomes more favourable for all frequencies when the segment length is increased.
Figure 3.2: Average $v_k$ (shape parameter of the Gamma PDF) over all unclassified segments for segment lengths $N_f = 5, 10, 20, 30$. 
Note that for the same second moment, the Exponential PDF has a heavier tail than the Rayleigh PDF’s; that is, the distribution tail becomes heavier as the segment length is increased. The fixed-$v$ ($v=8$) Gamma PDF is more favourable compared to the Exponential PDF when $N_f$ is low, but in general, Rayleigh PDF is still more favourable.

The percentages of segments with minimum $\chi^2$-values for $N_f = 20$ and 30 are plotted in Figure 3.4. The number of bins in the $\chi^2$-value calculation is set to 5, and the bin intervals are set such that for all bins, $p_i$’s are equal to $\frac{1}{5}$. The figures agree with the results obtained from the Log-Likelihood criteria, i.e. the Exponential PDF is preferred.

In the next section, the goodness-of-fit evaluation for the classified speech PDF is discussed.

### 3.3 The PDF of Classified Speech Spectral Amplitude

The speech sample used in this study is identical to that used in the unclassified study. Segmentation and labelling were performed manually. In the evaluation, we grouped the speech classes in Table 2.1 as follows. Vowels, glides, and diphthongs are analyzed as a group; fricative and affricates are also analyzed together; nasal is analyzed as its own class. The statistics of plosive sounds were not evaluated as their burst and aspiration durations were too short to enable accurate parameter estimation.

Firstly, let us describe the statistics of vowels, glides, and diphthongs. We picked 8 occurrences of each of these phonemes in Table 2.1. The average duration of all
3.3. THE PDF OF CLASSIFIED SPEECH SPECTRAL AMPLITUDE

Figure 3.3: Percentages of unclassified segments with maximum $L_{X_k}^{\text{Rayl}}$, $L_{X_k}^{\text{Exp}}$, and $L_{X_k}^{\gamma}$ values.

(a) $N_f = 5$

(b) $N_f = 10$
3.3. THE PDF OF CLASSIFIED SPEECH SPECTRAL AMPLITUDE

Figure 3.3: [cont.] Percentages of unclassified segments with maximum $\mathcal{L}_{X_k}^{\text{Rayl}}$, $\mathcal{L}_{X_k}^{\text{Exp}}$, and $\mathcal{L}_{X_k}^{\gamma}$ values.
Figure 3.4: Percentages of unclassified segments with minimum $\chi^2_{X_{x_k}}$, $\chi^2_{X_{y_k}}$, and $\chi^2_{X_{z_k}}$ values.
these segments is approximately 90 ms (approximately 10 frames). Segments that are less than 3 frames were not included. For each segment, the MLE parameters and the Log-Likelihood values of each candidate PDF were evaluated. To evaluate the Gamma PDF, a similar procedure is used as in the previous section—the average $v_k$’s over all segments in the sound group are evaluated, and a fixed $v_k$ is determined from these values. The average $v_k$ plot can be found in Figure 3.5; from this plot we fixed $v_k = 5$ for all $k$. The $\chi^2$-values were not calculated since the segments were too short in duration (< 20 frames). The percentages of segments with maximum $L_{X_k}^{\text{Rayl}}$, $L_{X_k}^{\text{Exp}}$, and $L_{X_k}^{\gamma}$ values are plotted in Figure 3.6. It suggests that the Gamma PDF with $v = 5$ represents the non-consonantal STSAs better than the other candidate PDFs. Also, the curve corresponding to the Gamma-PDF is closely followed by that corresponding to the Rayleigh PDF.

For nasal segments, we picked 10 occurrences of each phoneme /m, n, η/. The average segment duration of these segments is approximately 90 ms. The average $v_k$ of the nasal segments is plotted in Figure 3.7, and from this figure, we again chose a fixed $v_k = 5$ for all $k$. The percentages of segments with maximum $L_{X_k}^{\text{Rayl}}$, $L_{X_k}^{\text{Exp}}$, and $L_{X_k}^{\gamma}$ values are plotted in Figure 3.8. A similar trend to that observed in Figure 3.6 is obtained.

The evaluation of fricative segments is summarized in Figures 3.9 and 3.10. These segments have an average duration of approximately 90 ms. The trends of the percentage curves are similar to those in Figure 3.6 and 3.8. The Gamma or Rayleigh PDF is still preferable to the Exponential PDF.

These results did not indicate clear distinctive PDF models of different speech classes. For non-consonantal, and especially nasal classes, the Exponential PDF might
Figure 3.5: Average $v_k$ (shape parameter of the Gamma PDF) over the representative vowel, glide, and diphthong segments in the continuous speech sample.
Figure 3.6: Percentages of vowel, glide, and diphthong segments in the continuous speech sample with maximum $\mathcal{L}_{X_k}^{\text{Rayl}}$, $\mathcal{L}_{X_k}^{\text{Exp}}$, and $\mathcal{L}_{X_k}^{\gamma}$ values.
Figure 3.7: Average $v_k$ (shape parameter of the Gamma PDF) over the representative nasal segments in the speech sample.
Figure 3.8: Percentages of nasal segments in the speech sample with maximum $L_{X_k}^{\text{Rayl}}$, $L_{X_k}^{\text{Exp}}$, and $L_{X_k}^{\gamma}$ values.
Figure 3.9: Average $v_k$ (shape parameter of the Gamma PDF) over the representative fricative and affricate segments in the speech sample.
Figure 3.10: Percentages of fricative and affricate segments in the speech sample with maximum $\mathcal{L}_{X_k}^{\text{Rayl}}$, $\mathcal{L}_{X_k}^{\text{Exp}}$, and $\mathcal{L}_{X_k}^{\text{V}}$ values.
3.4 Summary and Conclusions

In this chapter, we investigated the parametric PDF models for the STSA of unclassified and classified speech. The model assumption is similar to that proposed in [16]; that is, the STSAs are independent in both time and frequency. An identical parametric PDF is assumed for all frequency bins and frames, but the PDF parameters are obtained using a weighted average of previous estimates. The study aimed to evaluate the fits of several parametric PDFs to the empirical STSA data.

The candidate PDFs are the Rayleigh, Exponential, and Gamma PDFs. Firstly, the parameters were obtained using the corresponding Maximum-Likelihood Estimators (MLEs). We varied the number of frames used to calculate these ML parameters. We also proposed to use the Log-Likelihood function, as well as the $\chi^2$ function, to rank the goodness-of-fit of the different candidate PDFs to the empirical data. A higher Log-Likelihood and a smaller $\chi^2$ values indicate that the assumed PDF fits the empirical data better.

For unclassified speech, we varied the number of frames over which the speech signal is assumed stationary. Using the evaluation method described above, we concluded that the Rayleigh PDF models shorter segments better. However, as the segment length increases, the Exponential PDF becomes preferable. For classified speech, we did not find clear model distinctions among different classes. In general, the Gamma PDF with shape parameter $v = 5$, is preferred, but it is closely followed...
by the Rayleigh PDF. Furthermore, for STSAs in the 1000-3000 Hz frequency range, the percentage of segments that are better modeled by the Exponential PDF is higher in the vowel, glide, and diphthong group, and especially in the nasal class, compared to that in the fricative class.

The results in this chapter suggest that Rayleigh and Gamma ($v = 5$) PDFs might model classified STSAs equally well. The STSA estimators that are based on these assumptions will be investigated in the next chapter.
Chapter 4

Noise Reduction Methods:
Literature Review and Proposed Methods

In this chapter, we discuss methods for noise reduction. In particular, the STFT-domain Bayesian noise reduction methods are presented in more detail. Section 4.1 provides an overview of noise reduction methods—a literature review of single-microphone noise reduction methods is presented. Section 4.2 provides a literature review of Bayesian noise reduction methods. A more detailed overview of Ephraim and Malah’s estimator [16] is presented in subsection 4.2.1. This estimator assumes a Rayleigh distribution for the speech STSA. Section 4.3 presents the proposed Gamma-PDF based spectral amplitude estimators. The subjective and objective performances of these estimators are then reported in Section 4.4. Finally, Section 4.5 provides the summary of this chapter.
4.1 Overview of Noise Reduction Methods

In the context of speech enhancement, noise reduction methods aim to improve the perceptual quality of noisy speech by removing the background noise. The topic can further be grouped into categories according to the problem settings, such as the availability of a noise reference, the number of available microphones, the noise types, and the way the noise interacts with the speech. Different signal processing techniques have been developed to deal with problems in different categories. For example, to enhance speech corrupted by reverberation, the signal is modeled as a convolution between the clean speech and the impulse response of the environment [4, Ch.11]. In the cocktail-party noise problem, the objective is to extract a speaker’s speech amongst multiple speakers in the background. In this case, the noisy signal is modeled as a linear combination of independent sources that are delayed and filtered [30].

As we have mentioned previously, a common procedure in non-stationary signal processing is to employ frame-based processing. In speech applications, the frame duration may vary from 16 to 32 ms. The appropriate frame duration that should be used in an application usually depends on several factors, some of which may conflict with one another. A longer frame length allows for finer frequency resolution—a benefit when a frequency-domain manipulation is performed, but decreases time resolution. Some speech sounds, such as plosives, are short in duration and characterized by the temporal envelope changes. On the other hand, vowel segments are usually longer and steadier. Variable-rate processing can also be employed, and it was claimed to significantly reduce the computational load [28].

The choice of signal representation is also a factor that should be considered. Noise
reduction methods employing the following signal representations have been proposed in the literature: Autoregressive coefficients (parametric model) [17, 14, 38], Short-Time Fourier Transform (STFT) [6, 5, 36, 16, 15, 34, 46, 44, 29], cepstrum [1], and signal subspace (Karhunen-Loeve Transform (KLT)) [18, 32, 9]. We chose the STFT representation because it allows for direct manipulations to the spectrum amplitude of the signal. It is reported that the KLT-domain algorithms have been more successful in reducing the musical residual noise—a common artifact produced by noise reduction methods [18].

We will discuss a noise reduction system that assumes an additive noise model as follows

\[ y(t) = x(t) + w(t), \]  

(4.1)

where \( y, x, \) and \( w \) denote noisy speech, clean speech, and noise signals, respectively. Furthermore, we assume that only one microphone is available, i.e., only the noisy speech is available as reference. A Voice Activity Detector (VAD) is also assumed to be available. This allows us to estimate and update the noise information during the time intervals when the VAD indicates silence.

Let us denote the STFT coefficients of the clean speech, noise, and noisy speech signals at frequency index \( k \) (from 1 to \( N \)) and frame index \( m \) respectively as follows,

\[ X_k(m) \triangleq X_k(m)e^{j\theta_{X,k}(m)}, \]

\[ W_k(m) \triangleq W_k(m)e^{j\theta_{W,k}(m)}, \]

\[ Y_k(m) \triangleq Y_k(m)e^{j\theta_{Y,k}(m)}. \]

An early proposed method for noise reduction in the spectral domain was the spectral subtraction estimator proposed by Boll in [6]. Assuming that the noise power
4.1. OVERVIEW OF NOISE REDUCTION METHODS

spectrum, $E[W_k^2(m)]$, is known, the clean speech STSA is estimated as follows,

$$\hat{X}_k(m) = \max \left\{ Y_k(m) - \sqrt{E[W_k^2(m)]}, 0 \right\}.$$  \hspace{1cm} (4.2)

This estimation is often referred to as the spectral magnitude subtraction. The complex STFT estimate is constructed by appending the noisy phase STFT to the estimated spectral amplitude, that is

$$\hat{\mathcal{D}}_k(m) = \hat{X}_k(m)e^{j\theta_{Y,k}(m)}.$$  \hspace{1cm} (4.3)

Other authors have also proposed estimators whose forms are similar to Boll’s estimator [5, 44]. Let us refer to this class of estimators as the spectral subtraction method. The general form of this estimator is

$$\hat{\mathcal{D}}_k(m) = \max\{(Y_k(m))^\alpha - \beta E[W_k(m)^\alpha])^{\frac{1}{2\alpha}}, \delta\} e^{j\theta_{Y,k}(m)},$$  \hspace{1cm} (4.4)

where $\alpha > 1$, $0 \leq \beta < 1$, and $\delta > 0$. When $\alpha = 2$ and $\beta = 1$, the method is often referred to as the power spectral subtraction method. For the spectral subtraction method in [6], the author reported an improvement in signal quality, specifically with respect to pleasantness and the amount of perceptible background noise. However, the Diagnostic Rhyme Test (DRT) showed that no intelligibility improvement is achieved.

A Maximum Likelihood spectrum amplitude estimator was proposed by McAulay and Malpass in [31]. In this case, the clean speech STFT is assumed constant, but unknown. The noise STFT is assumed to be a circularly symmetric zero-mean complex Gaussian r.v.; its real and imaginary parts have variances of $\frac{\lambda_w}{2}$. Furthermore, the STFT coefficients are independent in time and frequency. Given this model, the ML STSA estimator has the following expression:

$$\hat{X} = \frac{1}{2} \left( Y + \sqrt{Y^2 - \lambda_w} \right).$$  \hspace{1cm} (4.5)
The authors also proposed a two-state model, where the states correspond to speech is present and absent. Given the two-state model, the STSA estimator becomes

\[
\hat{X} = P(\text{speech is present}|Y)E[X|\text{speech is present}],
\]

where the last term in the right hand side of the equation is shown in (4.5). Similar to the spectral subtraction approach, the noisy phase is appended to \(\hat{X}\).

It is common for an STFT-domain noise reduction method (e.g., all STFT-domain methods cited above), to modify only the STFT magnitude and append the noisy phase. Nawab, et al. [35] showed that under some conditions, a perfect signal reconstruction can be achieved using magnitude-only representation. On the other hand, several studies, for example [36, 3], show that intelligibility can also be retained using a phase-only reconstruction (setting the magnitude to unity and using the clean STFT phase). To see the effect of phase in STFT reconstruction of a speech signal, Wang and Lim performed an experiment where they combined the STFT magnitude and phase of a speech sample with two different SNR levels [45]. Nine subjects were asked to compare a pair of a processed speech sample and the corresponding noisy (unprocessed) sample and choose the better-sounding one. An SNR equivalent is determined when the processed and unprocessed signals are chosen 50% of the time. Table 4.1 shows one of the results from the experiments. The STFT analysis and synthesis were performed using 512-pt DFT, Hanning window, and 10-kHz sampling rate. The study concluded that a more accurate phase representation does not bring any significant improvement in the equivalent SNR.

The advancement of noise reduction methods have been summarized in some publications, e.g., [26], [37], [14], [27], [4]. The majority of recent techniques are statistical-model based that employ Wiener filter [17, 14, 10], Bayesian filter [16, 29],
4.1. OVERVIEW OF NOISE REDUCTION METHODS

<table>
<thead>
<tr>
<th>Magnitude(dB)</th>
<th>0</th>
<th>5</th>
<th>15</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase (dB)</td>
<td>-25</td>
<td>-5</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>-25</td>
<td>-25</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
</tr>
<tr>
<td>-5</td>
<td>-25</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
</tr>
<tr>
<td>5</td>
<td>-25</td>
<td>-3.9</td>
<td>5</td>
<td>-5</td>
</tr>
<tr>
<td>15</td>
<td>-25</td>
<td>-3.9</td>
<td>4.9</td>
<td>15</td>
</tr>
<tr>
<td>25</td>
<td>-25</td>
<td>-4</td>
<td>6</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 4.1: Equivalent SNR (in dB) from Wang and Lim’s experiment [45].

or Kalman filter [38] principles. Statistical-model based noise reduction methods require the following assumptions [14]:

1. A joint statistical model of the speech and noise.
2. A distortion measure from which the estimator is derived.

Furthermore, to derive Bayesian filters, it is necessary to have an explicit expression of the joint distribution of speech and noise. An overview of the Bayes estimator, which is the approach used in our proposed noise reduction method, will be discussed in more detail in Section 4.2. In the following discussion, we briefly introduce the Wiener Filter as its performance is often compared to other proposed estimators’.

The Wiener filter, $g^\Delta = \begin{bmatrix} g_1 & g_2 & \cdots & g_L \end{bmatrix}^T$, is derived by minimizing the following Mean Square Error (MSE) function,

$$g : \arg \min_g E \left[ (x(m) - g^T y(m))^2 \right], \quad (4.7)$$

where $y(m)^\Delta = [y(m), y(m - 1), \cdots, y(m - L + 1)]^T$. $x$ and $y$ are clean and noisy speech signals in the corresponding operating domain. Furthermore, the speech and noise signals are assumed uncorrelated, and both are fully characterized by their autocorrelation matrices. Wiener filtering for speech enhancement has been applied
in the frequency domain [36], time domain [4, 10], and KLT domain [9]. In the frequency (STSA) domain, for the scalar case \((L = 1)\), the STSA estimator has the following expression (applies to each frequency bin and frame index),

\[
g = \frac{\lambda_X}{\lambda_X + \lambda_W} = \frac{\xi}{1 + \xi},\text{ where}
\]

\[
\xi \triangleq \frac{\lambda_X}{\lambda_W},
\]

where \(\lambda_X = E[X^2]\) and \(\lambda_W = E[W^2]\). \(\xi\) is referred to as the a-priori SNR. The complex STFT is again obtained by appending the noisy phase [36]. Moreover, since only noisy observations are available, \(\lambda_X\) and \(\lambda_W\) must be estimated from the noisy samples. The quality of the reconstructed signal is rather sensitive to these estimates [16].

The next section discusses Bayesian noise reduction methods. As we will see, the various Bayes estimators are developed based on different assumptions on the speech and noise PDFs, as well as the cost functions.

### 4.2 Overview of Bayesian Approach to Noise Reduction

In Bayes estimation, the unknown parameter of interest is assumed to be a random variable with a known a-priori PDF. Given the observation \(Y = y\), the Bayes estimator \(\hat{X}\) is derived to minimize the following Bayes risk:

\[
\hat{X} = \arg\min_{X^*} E[d(X, X^*)] = \arg\min_{X^*} \int d(X, X^*) f_{X|Y}(x|Y = y) dx,
\]

where \(d(., .)\) is the distance measure. For the MSE distance measure, \(d(X_1, X_2) = (X_1 - X_2)^2\), the corresponding **Minimum Mean Squared Error** (MMSE) estimator is
well-known to be the \textit{a-posteriori} mean, that is
\begin{equation}
\hat{X} = E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|Y = y) dx.
\end{equation}
(4.10)

Another Bayes estimator that we will also use is the Maximum A-Posteriori (MAP) estimator. The MAP estimator is derived to maximize the a-posteriori PDF of \(X\), as follows
\begin{equation}
\hat{X} = \arg \max_{x} f_{X|Y}(x|Y = y) = \arg \max_{x} f_{Y|X}(Y = y|x)f_{X}(x)dx.
\end{equation}
(4.11)

In this case, the distance measure is the “hit-or-miss” function \cite[pp.343-344]{*}, as follows
\begin{equation}
d(X_1, X_2) = \begin{cases} 
1, & \text{if } |X_1 - X_2| < \delta \\
0, & \text{otherwise} 
\end{cases},
\end{equation}
where \(\delta\) is positive and arbitrarily small.

Suppose that speech and noise signals, \(X\) and \(Y\) are independent. Given the explicit PDF expressions for the speech and noise, the a-posteriori PDF can be calculated using the Bayes’ rule as follows,
\begin{equation}
f_{X|Y}(x|Y = y) = \frac{f_{Y|X}(Y = y|x)f_{X}(x)}{\int_{-\infty}^{\infty} f_{Y|X}(Y = y|x)f_{X}(x)dx} = \frac{f_{W}(y - x)f_{X}(x)}{\int_{-\infty}^{\infty} f_{Y|X}(Y = y|x)f_{X}(x)dx}.
\end{equation}
(4.12)

Bayesian noise reduction methods can be categorized according to the approaches used to obtain \(f_X\). One of the simpler approaches assumes a parametric PDF for \(f_X\), and calculates the estimates of parameters frame-by-frame, as new observations arrive. The method was proposed in \cite{16}, and followed in \cite{15, 29, 8} by modifying the cost functions or parametric model assumptions. The other approach also uses a parametric PDF model, albeit a more complex one such as the mixture component.
model [34] or the Hidden Markov Model [17]. The model parameters are obtained during a training phase, and the training data should represent all observations in the test phase. Another category of noise reduction methods obtains the Bayes solution numerically from the training data, see for example, [40, 19, 46]. The STSA estimator used in the proposed noise reduction method utilizes the simpler frame-by-frame approach. A literature review of various STSA estimators that use this approach is presented in the following discussion.

In [16], Ephraim and Malah proposed a Bayes STSA estimator which models the speech and noise STFT components with complex circularly symmetric zero-mean Gaussian random variables. The speech and noise parameters, $\lambda_X(m)$ and $\lambda_W(m)$, are estimated every time a noisy frame arrives. The speech and noise signals are assumed to be independent, and their STFT components are also assumed to be independent in time and frequency. Using a similar problem formulation, other authors have proposed various estimators by choosing a different parametric PDF for the speech signal or using a different distance measure. Table 4.2 lists the references that present various Bayes STSA estimators, as well as the employed PDF assumptions and distance measures. Furthermore, subjective performance evaluations of these algorithms can be found in [29, 24].

In Chapter 3, we evaluated some PDF models for classified speech STSA and found that the Rayleigh or Gamma ($v = 5$) parametric PDF may model the empirical data equally well. In the following subsection, the Rayleigh-MMSE estimator proposed in [16] is summarized.
4.2. OVERVIEW OF BAYESIAN APPROACH TO NOISE REDUCTION

Table 4.2: PDF assumptions and distortion measures used in Bayesian speech estimators proposed in the literature

<table>
<thead>
<tr>
<th>PDF Models for Signal/Noise STFT Coefficients</th>
<th>Distortion Measure</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian/Gaussian</td>
<td>$d_{SE}(X, X) = (X - X)^2, \ p &gt; -2$</td>
<td>[16]</td>
</tr>
<tr>
<td></td>
<td>$d_{SE-LSA}(X, \hat{X}) = (\log X - \log \hat{X})^2$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d_{WSE}(X, \hat{X}) = (X - X)^2X^p$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d_{1S}(X, \hat{X}) = (X^2 - \hat{X}^2)^2$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d_{COSH}(X, \hat{X}) = \cosh(\log \frac{X}{\hat{X}}) = \frac{1}{2} \left[ \frac{X}{\hat{X}} + \frac{\hat{X}}{X} \right] - 1$</td>
<td>[29]</td>
</tr>
<tr>
<td></td>
<td>$d_{WLR}(X, \hat{X}) = (\log X - \log \hat{X})(X - \hat{X})$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d_{MIS}(X, \hat{X}) = \exp \left( X - \hat{X} \right) - (X - \hat{X}) - 1$</td>
<td></td>
</tr>
<tr>
<td>Laplacian/Gaussian</td>
<td>$d_{SE}(X, X) = (X - X)^2$</td>
<td>[8]</td>
</tr>
</tbody>
</table>

4.2.1 Rayleigh-MMSE Estimator [16]

The estimator supposes that the real and imaginary parts of speech STFT coefficients have the following distributions:

$$\Re X, \Im X \sim \mathcal{N}(0, \frac{\lambda_X}{2}),$$
$$\Re W, \Im W \sim \mathcal{N}(0, \frac{\lambda_W}{2}).$$  \hspace{1cm} (4.13)

For simplicity of notation, let us denote the complex-valued PDF by $X \sim \mathcal{N}_C(0, \lambda_X)$ and $W \sim \mathcal{N}_C(0, \lambda_W)$. The explicit expressions for the PDFs are

$$f_X(x) = \frac{1}{\pi \lambda_X} \exp \left( -\frac{|x|^2}{\lambda_X} \right),$$
$$f_W(w) = \frac{1}{\pi \lambda_W} \exp \left( -\frac{|w|^2}{\lambda_W} \right).$$
4.2. **OVERVIEW OF BAYESIAN APPROACH TO NOISE REDUCTION**

We assume that $\lambda_X$ and $\lambda_W$ are known. Furthermore, speech and noise are also assumed independent, so that

$$\mathcal{Y} \sim \mathcal{N}_C(0, \lambda_Y = \lambda_X + \lambda_W).$$

The conditional distribution of $Y$ given $X$ is the Rician PDF, which has the following expression

$$f_{Y|X}(y|x) = \frac{2y}{\lambda_W} e^{-\frac{(x^2+y^2)}{2\lambda_W}} I_0 \left( \frac{2xy}{\lambda_W} \right), \quad x \geq 0, y \geq 0$$ (4.14)

where

$$I_n(x) = \frac{1}{\pi} \int_0^{\pi} \cos(n\theta) e^{x \cos \theta} d\theta$$

denotes the $n$-th order modified Bessel function of the first kind. The speech STSA, $X$, has a Rayleigh distribution, that is

$$f_X(x) = \frac{2x}{\lambda_X} e^{-\frac{x^2}{2\lambda_X}}, \quad x \geq 0.$$ (4.15)

Using equations (4.14) and (4.15) to obtain the a-posteriori PDF expression (4.9) and substituting it into (4.10), we obtain the following expression for $\hat{X}$

$$X_{\text{Rayl}}^{\text{NMSE}} = \frac{\int_{0}^{\infty} x^2 e^{-\left(\frac{1}{\lambda_X} + \frac{1}{\lambda_W}\right)x^2} I_0 \left( \frac{2xy}{\lambda_W} \right) dx}{\int_{0}^{\infty} x e^{-\left(\frac{1}{\lambda_X} + \frac{1}{\lambda_W}\right)x^2} I_0 \left( \frac{2xy}{\lambda_W} \right) dx}$$

$$= \Gamma(1.5) \frac{\sqrt{\nu}}{\gamma} e^{-\frac{\nu}{2}} \left[ (1 + \nu) I_0 \left( \frac{\nu}{2} \right) + \nu I_1 \left( \frac{\nu}{2} \right) \right] Y,$$ (4.16)

where

$$\xi \triangleq \frac{\lambda_X}{\lambda_W} \quad \text{(a-priori SNR)},$$

$$\gamma \triangleq \frac{\lambda^2}{\lambda_W} \quad \text{(a-posteriori SNR)},$$

$$\nu \triangleq \frac{\xi}{1+\xi}\gamma.$$ (4.17)
4.2. OVERVIEW OF BAYESIAN APPROACH TO NOISE REDUCTION

Let us define the estimator gain to be

\[ G = \frac{\hat{X}}{Y}, \]

and denote \( G_{\text{Rayl}}^{\text{MMSE}} \) to be the corresponding gain function of the estimator in (4.16). The \( G_{\text{MMSE}} \) as functions of the a-priori and a-posteriori SNR are plotted in Figures 4.1 and 4.2. The noise variance, \( \lambda_W \) is set to 1. We noted that:

1. The gain is an increasing function with respect to the a-priori SNR.
2. The gain is a decreasing function with respect to the a-posteriori SNR.

Furthermore, strong attenuation is only applied to samples with low a-priori SNR. Also, for \( \gamma - 1 = 20 \text{ dB} \), the gain curve almost coincides with the Wiener gain. The fact that \( G_{\text{MMSE}}^{\text{Rayl}} \approx \frac{\xi}{1 + \xi} \) if the a-posteriori SNR is high has also been shown analytically in [16]. Therefore, if the a-posteriori SNR is high, the a-priori SNR is the dominant factor that determines the suppression level.

To obtain the a-priori SNR estimate, \( \hat{\xi}(m) \), Ephraim and Malah proposed the following decision-directed estimator,

\[ \hat{\xi}(m) = \alpha_\xi G^2(\xi(m - 1), \gamma(m - 1)) \gamma(m - 1) + (1 - \alpha_\xi)\max\{\gamma(m) - 1, 0\}, \quad (4.18) \]

where \( 0 < \alpha_\xi < 1 \) and \( \alpha_\xi \) is usually close to 1 (in the range of .9-.995). The decision-directed estimator has been reported to considerably reduce the amount of perceptible musical noise [16, Section VI, Case 3], [7]. The closer is \( \alpha_\xi \) to one, the lesser is the audible musical noise. However, the speech itself becomes more distorted. Furthermore, it is also a common practice to impose a minimum threshold on \( \hat{\xi}_k(m) \), that is

\[ \hat{\xi}(m) = \max \{\delta, \alpha_\xi G^2(\xi(m - 1), \gamma(m - 1)) \gamma(m - 1) + (1 - \alpha_\xi)\max\{\gamma(m) - 1, 0\}\}, \quad (4.19) \]
4.2. OVERVIEW OF BAYESIAN APPROACH TO NOISE REDUCTION

Figure 4.1: Gain function of the Rayleigh-MMSE estimator as a function of $\xi$.

Figure 4.2: Gain function of Rayleigh-MMSE estimator as a function of $\gamma$. 

\[ 20 \log_{10} G_{\text{RaylMMSE}}(\xi) \]

\[ \lambda_X = -5 \, \text{dB} \]
\[ \lambda_X = 0 \, \text{dB} \]
\[ \lambda_X = 5 \, \text{dB} \]
\[ \lambda_X = 10 \, \text{dB} \]
where $\delta > 0$ acts as a spectral floor.

Let us define the true a-priori SNR value by $\xi^*$ and its perturbed value by $\tilde{\xi} = \xi^* + \Delta\xi$. Let us define the normalized residual MSE and the normalized bias to be

\[
\epsilon(\xi^*, \tilde{\xi}) \doteq \frac{E[(X - \hat{X})^2]}{E[(X - E[X])^2]}, \quad \text{and}
\]
\[
B(\xi^*, \tilde{\xi}) \doteq \frac{E[X - \hat{X}]}{E[X]},
\]

respectively. The analytical expressions for both $\epsilon(\xi^*, \tilde{\xi})$ and $B(\xi^*, \tilde{\xi})$ when $\hat{X}_{\text{Rayl}}^{\text{MMSE}}$ and $\hat{X}_{\text{Wiener}}$ are calculated using the perturbed $\xi$ value are derived in [16]. It is shown that both estimators are less sensitive to an overestimation of $\xi$, compared to an underestimation, in terms of the normalized residual MSE and bias, therefore justifying the use of a spectral floor for the estimation of $\xi$.

The next section presents the derivation of the proposed MAP and MMSE estimators which assume the Gamma distribution for the speech STSA.

### 4.3 Proposed Gamma-PDF Based Estimators

Motivated by the conclusions that we obtained in Chapter 3, we derived some spectrum amplitude estimators assuming the Gamma PDF for the speech STFT magnitude. The STFT phase is assumed to be independent of the STFT magnitude and has a Uniform distribution. Let us denote the PDFs as follows

\[
f_X(x) = \frac{x^{v-1}}{\Gamma(v)\zeta^v}e^{-\frac{x}{\zeta}}, \quad x \geq 0 \quad (4.20)
\]
\[
f_{\theta_X}(\theta) = \frac{1}{2\pi}, \quad 0 \leq \theta \leq 2\pi \quad (4.21)
\]
4.3. PROPOSED GAMMA-PDF BASED ESTIMATORS

Furthermore, the noise $\mathcal{W}$ is assumed to be a circularly symmetric zero-mean complex Gaussian random variable (4.13). Subsections 4.3.1 and 4.3.2 present the derivation of the MAP and MMSE estimators, respectively.

### 4.3.1 Gamma-MAP Estimator

The MAP estimator is obtained by maximizing the a-posteriori probability of $X$, which is also equivalent to maximizing the following function

$$
\hat{X}_{\text{MAP}} = \arg \max_X f_{Y|X}(y|x) f_X(x).
$$

(4.22)

Equivalently, $\hat{X}_{\text{MAP}}$ is the value $X$ that solves the following equation

$$
\frac{\partial}{\partial x} \left[ \log f_{Y|X}(y|x) + \log f_X(x) \right] = 0.
$$

(4.23)

Recall that $f_{Y|X}$ is the Rician distribution as shown in (4.14). Assuming that $\frac{2XY}{\lambda_W} \gg 0$ (i.e., high-SNR), the following approximation holds for $I_\alpha(z)$,

$$
I_\alpha(z) \approx \frac{1}{\sqrt{2\pi z}} e^z, \alpha << |z|
$$

(4.24)

Substituting equation (4.14), (4.20), and (4.24) into (4.23), we have

$$
X^2 - \left( Y - \frac{\lambda_W}{2\zeta} \right) x - \frac{\lambda_W}{2} \left( v - \frac{3}{2} \right) = 0
$$

(4.25)

Solving for $x$ gives the minimum point in the quadratic equation

$$
\hat{X}_{\text{MAP}}^\gamma = \frac{Y}{2} - \frac{\lambda_W}{4\zeta} + \sqrt{\left( \frac{Y}{2} - \frac{\lambda_W}{4\zeta} \right)^2 + \frac{\lambda_W}{2} \left( v - \frac{3}{2} \right)}
$$

(4.26)

Let us denote the second moment of $X$ as follows,

$$
\lambda_X^2 \overset{\Delta}{=} E \left[ X^2 \right] = v(v+1)\zeta^2.
$$
Writing the gain $G_{\gamma}^{\text{MAP}} = \frac{\hat{X}_{\gamma}^{\text{MAP}}}{Y}$ in terms of $\xi = \frac{\Delta_{X}}{\lambda_{W}}$ and $\gamma = \frac{\Delta_{Y}}{\lambda_{W}}$, we have the following expression,

$$
\hat{G}_{\gamma}^{\text{MAP}}(\xi, \gamma) = \frac{1}{2} - \frac{\sqrt{v(v+1)}}{4\sqrt{\xi}\gamma} + \sqrt{\left(\frac{1}{2} - \frac{\sqrt{v(v+1)}}{4\sqrt{\xi}\gamma}\right)^2 + \frac{v - \frac{3}{2}}{2\gamma}}. \tag{4.27}
$$

The plots of the gain function $G_{\gamma}^{\text{MAP}}$ for $v = 5$ are plotted in Figures 4.3 and 4.4. Figure 4.3 shows the gain as a function of the a-posteriori (or instantaneous) SNR. The $\xi$’s are set so that $\lambda_{X} = -5, 0, 5, \text{ and } 10 \text{ dB}$, and $\lambda_{W}$ is set to 1. Figure 4.4 shows the gain as a function of the a-priori SNR ($\xi$). From these figures we observed the following:

1. The gain is an increasing function of $\xi$.

2. The gain approaches unity when the a-posteriori (instantaneous) SNR is high (> 30 dB).

3. For a constant $\xi$, the gain decreases as $\gamma$ increases, reaches a minimum, and increases again to unity.

Furthermore, Figure 4.4 shows that the gain curve keeps decreasing until it approaches the Wiener gain (dashed curve) as $\gamma$ is increased, and moves away from it (increases again) as $\gamma$ increases.

To obtain the gain parameter, $\xi(m)$, we employed the decision-directed estimator (4.19) as follows,

$$
\hat{\xi}(m) = \max \left\{ \alpha[G_{\gamma}^{\text{MAP}}(m-1)]^2 + (1 - \alpha) \max [\gamma(m) - 1, 0], \delta \right\}. \tag{4.28}
$$

The performance of the Gamma-MAP estimator is reported in Section 4.4. In the next subsection, the Gamma-MMSE estimator is discussed.
4.3. PROPOSED GAMMA-PDF BASED ESTIMATORS

Figure 4.3: Gain function of the Gamma-MAP estimator as a function of the a-posteriori SNR.
4.3. PROPOSED GAMMA-PDF BASED ESTIMATORS

Figure 4.4: Gain function of the Gamma-MAP estimator as a function of the a-priori SNR.
4.3. PROPOSED GAMMA-PDF BASED ESTIMATORS

4.3.2 Gamma-MMSE Estimator

Recall that under the MMSE condition, the Bayes estimate is the a-posteriori mean, i.e.,

\[
\hat{X} = \int_0^\infty x f_{Y|X}(y|x) f_X(x) dx / \int_0^\infty f_{Y|X}(y|x) f_X(x) dx.
\]

Substituting the Rician PDF (4.14) into \( f_{Y|X}(y|x) \), and the Gamma distribution (4.20) into \( f_X(x) \) into the above equation gives the following expression,

\[
\hat{X}_{\text{MMSE}}^\gamma = \int_0^\infty x^v e^{-\frac{x^2}{2\lambda W}} I_0 \left( \frac{2xy}{\sqrt{\lambda W}} \right) dx / \int_0^\infty x^{v-1} e^{-\frac{x^2}{2\lambda W}} I_0 \left( \frac{2xy}{\sqrt{\lambda W}} \right) dx.
\]

(4.29)

We were not able to find the closed-form solution to the above equation. Replacing \( I_0 \) with its high SNR approximation for \( I_0 \) (4.24), the numerator and denominator in (4.29) have the following form

\[
\int_0^\infty x^{\alpha-1} e^{-\beta x^2-\gamma x} dx = (2\beta)^{-\frac{\alpha}{2}} \Gamma(\alpha) e^{\frac{\gamma^2}{4\beta}} D_{-\alpha} \left( \frac{\gamma}{\sqrt{2\beta}} \right), \quad \text{Re}\beta > 0, \text{Re}\alpha > 0,
\]

(4.30)

where the equality is given in [20, p.365, eq. 3.462.1]. \( D_\alpha(x) \) is the parabolic cylinder function [20, p.1028, eq.9.240], which is equal to

\[
D_\alpha(z) = 2^{\frac{\alpha}{2}} e^{-\frac{z^2}{4}} \left\{ \frac{\sqrt{\pi}}{\Gamma(\frac{1-\alpha}{2})} \Phi \left( -\frac{\alpha}{2}, \frac{1}{2}; \frac{z^2}{2} \right) - \frac{\sqrt{2\pi}}{\Gamma(\frac{\alpha}{2})} \Phi \left( \frac{1-\alpha}{2}, \frac{3}{2}; \frac{z^2}{2} \right) \right\},
\]

(4.31)

where \( \Phi(a, b; z) \) denotes the confluent hypergeometric function [20, Sec. 9.2, p.1022], that is

\[
\Phi(a, b; z) = 1 + \frac{a}{b} z + \frac{a(a+1)}{b(b+1)} \frac{z^2}{2!} + \frac{a(a+1)(a+2)}{b(b+1)(b+2)} \frac{z^3}{3!} + \cdots.
\]

(4.32)

The approximated solution of \( \hat{X}_{\text{MMSE}}^\gamma \) is therefore,

\[
\hat{X}_{\text{MMSE}}^\gamma = \sqrt{\frac{\lambda W}{2}} \left( v - \frac{1}{2} \right) \frac{D_{-(v+0.5)}(z)}{D_{-(v-0.5)}(z)},
\]

(4.33)
where \( z = \sqrt{\frac{\lambda_W}{2}} \left( 1 - \frac{2Y}{\lambda_W} \right) \). Writing the gain expression, \( G_{\text{MMSE}}^\gamma = \frac{\hat{X}_{\text{MMSE}}}{Y} \), in terms of \( \xi \), and \( \gamma \) gives the following expression,

\[
\begin{align*}
    G_{\text{MMSE}}^\gamma(\xi, \gamma) &= \frac{(v - \frac{1}{2}) D_{-(v+0.5)}(z) - (v + 0.5)(z)}{\sqrt{2\gamma} D_{-(v-0.5)}(z)}.
\end{align*}
\]

The plots of the Gamma-MMSE suppression curves for \( v = 5 \) are shown in Figures 4.5 and 4.6. The confluent hypergeometric function was approximated by summing the first 500 terms in the series. Figure 4.5 plots the gain as a function of \( \gamma \). The \( \zeta \)'s are set so that \( \xi = -5, 0, 5, \) and \( 15 \) dB, and the noise variance \( \lambda_W \) is set to 1. Figure 4.6 shows the gain as a function of \( \xi \). These figures show similar trends as those observed in the Gamma-MAP gain, i.e., the gain is an increasing function of \( \xi \), and for a constant \( \xi \), the gain decreases as \( \gamma \) increases, approaches a minimum point, which is the Wiener gain, and increases again. Moreover, all of the gain curves in Figure 4.5 converge to unity gain when the a-posteriori SNR is high (\( \zeta = 35 \) dB).

The gain values for \( \xi \) outside -5 and 25 dB range, and \( \gamma - 1 \) outside -10 and 25 dB range were not shown in the plots since numerical error in the function calculation has occurred for SNR below this range (the function \( D_{-\alpha}(z), \alpha > 0 \) approaches \( \infty \) if the SNR is too low). Recall that we have approximated \( I_0 \) with its high-SNR approximation (4.24). Furthermore, there is also numerical error caused by truncating the confluent hypergeometric terms.

Comparing the Gamma (MAP and MMSE) estimators with the Rayleigh-MMSE estimator, we noted that the Gamma-based estimators give more consideration to the a-posteriori SNR information compared to that given by the Rayleigh-MMSE estimator. In the Rayleigh-MMSE estimator, the suppression level is a function of
4.3. PROPOSED GAMMA-PDF BASED ESTIMATORS

Figure 4.5: Gain function of the Gamma-MMSE estimator as a function of the a-posteriori SNR.
4.3. PROPOSED GAMMA-PDF BASED ESTIMATORS

Figure 4.6: Gain function of Gamma-MMSE estimator as a function of the a-priori SNR.
the a-priori SNR only if the a-posteriori SNR exceeds a certain level. In contrast, the Gamma-MAP gain eventually reaches unity if the a-posteriori SNR level is high enough, regardless of the a-priori SNR value. In the Gamma-MMSE estimator, for high a-priori SNR, the gain decreases and reaches the Wiener gain as the a-posteriori SNR increases. However, when the a-priori SNR is low, the gain decreases to the Wiener gain, but increases again as the a-posteriori SNR increases.

In the next section, performance evaluations of the three STSA estimators are reported.

4.4 Performance Comparison

4.4.1 Subjective Evaluation

Test Materials, Implementation, and Evaluation Methodology

The test materials used to evaluate the speech enhancement algorithms were the same as that collected for the PDF study in Chapter 3 (an approximately one-and-a-half minute read speech). The noise signal was scaled according to the desired SNR and added to the clean sample. The noise types were white, car, and babble noise. White noise samples were generated with the MATLAB function \texttt{randn}, while the car and babble noise samples were obtained from the Signal Processing Information Database website: \texttt{http://spib.rice.edu/spib/signal.html}. The following SNR definition was used to obtain the appropriate noise power level:

\[
\text{SNR} = \frac{\text{85\%-percentile of the power of non-silence frames in clean speech}}{\text{mean power of noise frames}}. \quad (4.35)
\]
The simulation was performed with the following STFT analysis and synthesis parameters: 32-ms normalized Hanning window (3.11), 75% frame overlap, and \( N = 512 \) point FFT. The noise power spectrum, \( \lambda_W(m) \), was estimated and updated when the VAD indicated silence.

The VAD’s decision was based on two variables: the frame SNR threshold, \( \theta_N \), and the number of previous consecutive frames whose SNR are less than a specified threshold, \( T_{\text{Silence}} \). A frame is indicated as silence if the number of previous frames whose SNR are less than \( \theta_N = 3 \) dB exceeds \( T_{\text{Silence}} = 8 \) frames (100 ms). When the VAD indicates silence, the noise variance is updated as follows,

\[
\lambda_W(m) = \alpha_W \lambda_W(m-1) + (1 - \alpha_W)Y_k^2(m),
\]

(4.36)

where \( \lambda_W < 1 \) and is close to 1. Furthermore, all estimators employed the decision-directed estimator, and the smoothing parameters, \( \alpha_\xi \) and \( \alpha_W \), were tuned according to the noise type (see Table 4.3).

We selected an SNR level of 0 dB. After noise addition, the test materials were processed by four noise reduction algorithms: Wiener, Rayleigh-MMSE, Gamma-MAP, and Gamma-MMSE estimators. A sentence from the speech was then presented to the subjects, and they were asked to evaluate the quality of the noise reduced speech. In total, there were 5 subjects; all of them were between 20 to 30 years of age and had self-reported normal hearing. All, but one, were non-native English speakers but with fluent knowledge of the language.

The evaluation was performed according to the methodology described in ITU-T P.835 (Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms [24]). Subjects were asked to rate the signal on the discrete scale from 1 (worst) to 5 (best) in terms of:
1. Speech signal distortion (SIG). The five-point scale corresponds to:

<table>
<thead>
<tr>
<th>Rating</th>
<th>Quality of Signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Very natural, no degradation</td>
</tr>
<tr>
<td>4</td>
<td>Fairly natural, little degradation</td>
</tr>
<tr>
<td>3</td>
<td>Somewhat natural, somewhat degraded</td>
</tr>
<tr>
<td>2</td>
<td>Fairly unnatural, fairly degraded</td>
</tr>
<tr>
<td>1</td>
<td>Very unnatural, very degraded</td>
</tr>
</tbody>
</table>

2. Background noise intrusiveness (BKG). The five-point scale corresponds to:

<table>
<thead>
<tr>
<th>Rating</th>
<th>Quality of Background Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Not noticeable</td>
</tr>
<tr>
<td>4</td>
<td>Somewhat noticeable</td>
</tr>
<tr>
<td>3</td>
<td>Noticeable but not intrusive</td>
</tr>
<tr>
<td>2</td>
<td>Fairly conspicuous, somewhat intrusive</td>
</tr>
<tr>
<td>1</td>
<td>Very conspicuous, very intrusive</td>
</tr>
</tbody>
</table>

3. The overall signal quality (OVL) using the Mean Opinion Score rating, that is:

1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent.

Each sentence was played three times, and at each time, the subject was instructed to rate one of the SIG, BKG, and OVL criteria. In total, there were 24 sentences, which correspond to 3 types of noise, 4 noise reduction algorithms, and 2 sentences for each noise and algorithm combination. Prior to testing, we presented the subjects with 12 training sentences (one example for each condition).
In calculating the Gamma-MMSE gain, the $\xi(m)$ and $\gamma(m)$ were limited to -5 and 25 dB range. Recall that (4.34) was obtained by assuming a high-SNR approximation. Furthermore, numerical error also occurred for $\xi$ and $\gamma$ greater than 25 dB, due to the series truncation in the confluent hypergeometric function (4.32). Furthermore, for the Gamma-based estimators, $v = 5$ was fixed as a constant.

For each signal condition, an informal listening test was first conducted to determine the smoothing parameters. The values that gave the best result were used in the experiment. Table 4.3 shows the smoothing parameters, $\alpha_\xi$ and $\alpha_W$, used in each signal condition.

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>Wiener</th>
<th>Rayleigh-MMSE</th>
<th>Gamma-MAP</th>
<th>Gamma-MMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>White + Car</td>
<td>$\alpha_\xi$</td>
<td>$\alpha_W$</td>
<td>$\alpha_\xi$</td>
<td>$\alpha_W$</td>
</tr>
<tr>
<td>Babble</td>
<td>.99</td>
<td>.9</td>
<td>.99</td>
<td>.9</td>
</tr>
</tbody>
</table>

**Results**

Figure 4.8 illustrates the waveforms of a noise-reduced speech sample processed by the four algorithms. From the figure, we can observe that the Gamma-MMSE estimator suppressed the least amount of noise. This is also reflected in the listening test. The mean SIG, BKG, and OVL scores for each signal condition can be seen in Figure 4.7. The results indicated that the estimators’ performances are dependent on the noise type. The estimators, notably the Wiener and Gamma-MAP filters, performed relatively well for car noise. Both estimators produce enhanced signals that sounded very similar to the clean speech. Babble noise is the hardest type of noise to be dealt with, reflected by the low scores. For this type of noise, the Rayleigh-MMSE estimator
4.4. PERFORMANCE COMPARISON

Figure 4.7: The mean SIG, BKG, and OVL scores obtained from the listening test.

seems to outperform the other algorithms. Moreover, the Wiener suppression gain provides the strongest suppression. This was reflected in the listening test where the residual noise sounded least noticeable, but the signal itself was more distorted.

4.4.2 Objective Evaluation

The test material used is identical to that used in the subjective evaluation test. In the following, the amounts of speech distortion and noise suppression caused by each estimator at various SNR levels and noise types are evaluated.

To quantify the amounts of signal distortion and noise reduction caused by each
4.4. PERFORMANCE COMPARISON

Figure 4.8: Waveforms of 0-dB white-noise corrupted speech processed by Wiener filter (third panel), Rayleigh-MMSE estimator (fourth panel), Gamma-MAP estimator (fifth panel), and Gamma-MMSE estimator (sixth panel).
signal estimator, we propose to use the following procedure. The enhanced speech STFT was obtained as follows,

$$\hat{\mathcal{X}} = G(Y)\hat{\mathcal{Y}}$$

$$= G(Y)\mathcal{X} + G(Y)\mathcal{W}$$

Let us introduce the following notation,

$$\tilde{\mathcal{X}} = G\mathcal{X},$$

$$\tilde{\mathcal{W}} = G\mathcal{W},$$

where we have omitted the frequency- and frame-index notations for the sake of brevity. Let us also denote the corresponding time-domain signals by $\tilde{x}$ and $\tilde{w}$, respectively. They were obtained by performing an inverse STFT on $\tilde{\mathcal{X}}$ and $\tilde{\mathcal{W}}$, followed by the overlap-and-add operation.

To quantify the signal distortion, we propose to use the following measure,

$$v_{SD} = \Delta \sum_n (x(n) - \tilde{x}(n))^2 \sum_n x^2(n),$$

(4.37)

and to quantify the noise reduction factor, we propose to use the following measure,

$$v_{NR} = \Delta \sum_n w^2(n) \sum_n \tilde{w}^2(n).$$

(4.38)

A small $v_{SD}$ indicates a better signal quality while a large $v_{NR}$ indicates a better noise reduction.

The signal distortion and noise reduction factors for the four estimators and three noise types are plotted as a function of SNR in Figure 4.9. From the figure, we can see the following: 1) Both $v_{SD}$ and $v_{NR}$ for babble noise are by far the worst compared to the other two noise types, 2) Wiener filter, followed by the Gamma-MAP, Rayleigh-MMSE, and Gamma-MMSE filters, incurs the most to the least signal distortion and noise suppression, 3) For car noise, $v_{SD}$ approaches zero more quickly, with respect to increasing SNR, than for the other noise conditions.
4.4. PERFORMANCE COMPARISON

Figure 4.9: The signal distortion values (left column) and noise reduction factors (right column) of the enhanced speech samples as a function of SNR.
4.5 Summary and Conclusions

In this chapter, we have presented the background and the details of the spectral amplitude estimation block. In Section 4.1, a literature review of some popular noise reduction algorithms was presented, and we also introduced in more detail a kind of Bayesian algorithm which operates in the STFT domain. The signal model and assumptions were discussed, and various estimators derived by varying the a-priori PDF assumptions on the speech STSA, as well as the cost functions, were summarized in Table 4.2. A literature review of the Rayleigh-MMSE estimator [16], which is based on the Rayleigh PDF assumption and derived to minimize the Bayesian MSE cost function, was presented in subsection 4.2.1.

In Chapter 3, we concluded that the Rayleigh or Gamma PDF with a shape parameter, \( v = 5 \), are more favourable (with respect to the Log-Likelihood (3.9) and \( \chi^2 \) (3.10) criteria) to model the a-priori PDF of clean speech STSA. Therefore, we derived the Gamma-MAP and Gamma-MMSE estimators. Similar to other STFT estimators of this kind, the noisy phase is appended to the estimated STSA to create the complex STFT of the enhanced speech. The derivations as well as gain properties of these estimators were studied and compared to the Wiener and Rayleigh-MMSE estimators in Section 4.3. The Bayes estimators (Rayleigh-MMSE, Gamma-MAP, and Gamma-MMSE estimators) are functions of the a-priori SNR (\( \xi \)), as well as the a-posteriori SNR (\( \gamma \)). They are both increasing functions of \( \xi \). The Rayleigh-MMSE is a decreasing function of \( \gamma \), and its gain approaches the Wiener filter gain as \( \gamma \) is increased. The Gamma-MAP estimator behaves quite similarly when the a-posteriori SNR is low–increasing \( \gamma \) results in smaller gain. However, when \( \gamma \) is large enough (see Figure 4.3), the gain increases with increasing \( \gamma \), i.e., the Gamma-MAP
estimator weighs the a-posteriori SNR information more heavily compared to the other estimators. A similar trend was also observed in the Gamma-MMSE estimator.

The subjective and objective performance evaluations of the Wiener, Rayleigh-MMSE, Gamma-MAP, and Gamma-MMSE estimators were presented in Section 4.4. The test signals used in the subjective evaluation test consist of speech samples corrupted by white, car, and babble noise at 0 dB SNR. The evaluation procedure followed the ITU-T P.835 methodology, and the results can be seen in Figure 4.7. The objective evaluation was described in terms of signal distortion index (4.37) and noise reduction factor (4.38). The results can be seen in Figure 4.9. The Wiener estimator, followed by the Gamma-MAP, Rayleigh-MMSE, and Gamma-MMSE estimators incurred the most to the least amount of signal distortion as well as noise suppression.

In conclusion, it is not straightforward to determine which estimator to use if it needs to work well in all conditions. The decision depends on the tradeoff between signal distortion and noise reduction, but especially for nonlinear estimators, it is quite hard to analyze them theoretically. In terms of signal distortion/noise suppression, the proposed Gamma-MAP estimator is quite competitive to the Rayleigh-MMSE estimator. Moreover, the former has a simpler expression, and therefore, is more easily implementable.

The next chapter discusses the implementation details, as well as the performance evaluation, of the proposed class-based noise reduction system.
Chapter 5

Implementation of the Proposed System

This chapter discusses the implementation and simulation results of the proposed class-based noise reduction system. Section 5.1 discusses the implementation details of the cue enhancement strategies presented in Chapter 2. In Section 5.2, the performance of the proposed system was evaluated and compared to that of the Rayleigh-MMSE estimator. In this case, performance was measured in terms of the intelligibility of the output speech. Finally, the summary and conclusions of this chapter are presented in Section 5.3.

5.1 System Implementation

The proposed class-based noise reduction method has been presented in Chapter 2.3, and the summary can be found in Figure 2.5.

For plosives and fricatives, we aimed to enhance the place of articulation cue by...
5.1. **SYSTEM IMPLEMENTATION**

first analyzing the burst and aspiration (noisy) frames to identify the gross spectrum trend, and applying a spectral shaping gain to the enhanced STFT. Furthermore, to enhance the characteristics of plosive sounds, burst and aspirated frames are amplified by +6 and +3 dB, respectively. Fricative and nasal frames are also amplified by +3 dB. The amplification is performed after the core noise reduction block.

To estimate the gross spectrum amplitude trend in burst and aspiration frames, we perform the following procedure. We first find the frequency region with the highest energy concentration using a fourth-order Linear Prediction (LP) analysis [41, p.207]. The center frequency of this region is determined to be one of the two zeroes of the LP polynomial that is within 1-5 kHz frequency range. If both of them are located within 1-5 kHz, we choose the one that corresponds to a larger spectrum amplitude, i.e., choose \( \omega_c \) such that

\[
\omega_c = \arg \max_{\omega_1, \omega_2} \left| \frac{1}{A(e^{j\omega})} \right|,
\]

where \( A(e^{j\omega}) \) denotes the LP polynomial, and \( \omega_1, \omega_2 \) denote the zeroes of \( A(e^{j\omega}) \). If none of the zeroes lies within the 1-5 kHz range, no spectral shaping is performed after the noise suppression.

The spectral shaping for the burst and aspiration frames is performed by multiplying the enhanced STSA with a raised cosine window centered at frequency-bin \( k_c = \left\lfloor \frac{\omega_c}{2\pi} N \right\rfloor \). The raised cosine window has the following expression:

\[
H_k = a + \frac{b}{2} \left( 1 + \cos \left( \frac{1}{N\beta} (k - k_c) \right) \right), \quad k = 1 \cdots N,
\]

where \( a = 1 \) \((a \geq 1)\), \( b = 0.3 \) \((b > 0)\), and \( \beta \), which controls the lobe width, is set so that the lobe width is approximately 3000 Hz, that is,

\[
\beta = \frac{\text{Lobe width of the raised cosine window (in number of bins)}}{N}.
\]
5.1. SYSTEM IMPLEMENTATION

For fricative frames, the noisy frame is analyzed to determine if it has a highpass or flat trend by comparing the total power in the high-frequency region (4.5-8 kHz) to the total power in the lower region. A frame is determined as a highpass if it satisfies

\[ R = \frac{\sum_{k \in 4.5-8kHz} Y_k^2}{\sum_{k} Y_k^2} \geq \Delta_F, \]

where \( \Delta_F \) is a pre-specified threshold. If a highpass trend is indicated, the enhanced STSA is multiplied with a highpass filter with a magnitude frequency response shown in Figure 5.1. Otherwise, no spectrum shaping is performed.

In Chapter 2, we have also proposed to investigate the use of different STSA estimators for processing different speech classes. In Chapter 3, we concluded that the Gamma PDF, with \( v = 5 \), and the Rayleigh PDF are almost equally favoured as a parametric model for the PDF of classified speech STSA. In Chapter 4, we evaluated the Rayleigh-MMSE and Gamma-based estimators. The simulation results suggested
that compared to the Gamma-MAP estimator, the Rayleigh-MMSE estimator suppresses more of the background noise, but with a tradeoff of a higher signal distortion. Based on this result, we proposed to perform the following:

- For frames that were classified as vowels, diphthongs, glides, or nasals, we applied the Gamma-MAP estimator. Intuitively, these sound classes can afford more signal distortion, since they tend to be more steady.

- For frames that were classified as fricatives, affricates, and burst/aspiration in plosives, we applied the Rayleigh-MMSE estimator. Stop was processed using the Gamma-MAP estimator (stronger noise suppression).

This system was simulated, setting the smoothing factor in the a-priori SNR estimation, $\alpha_\xi$ to be .99. The output speech, however, sounded quite unnatural since the characteristics of speech segments processed with the different estimators sounded very different; those processed with the Rayleigh-MMSE estimator contained much more residual noise. Hence, in the intelligibility evaluation, we chose the same type of estimator to process all sound classes—in this case, we used the Rayleigh-MMSE estimator.

The proposed class-based noise reduction system is summarized in Figure 5.2. In the next section, we compare the performance of this system to that of the Rayleigh-MMSE estimator in terms of the intelligibility of the output speech.
Table 5.1: Lists 11 and 13 of the phonetically-balanced word list in [13, Appendix].

<table>
<thead>
<tr>
<th>List 11</th>
<th>List 13</th>
</tr>
</thead>
<tbody>
<tr>
<td>arc</td>
<td>bat</td>
</tr>
<tr>
<td>doubt</td>
<td>few</td>
</tr>
<tr>
<td>jab</td>
<td>jig</td>
</tr>
<tr>
<td>pond</td>
<td>nip</td>
</tr>
<tr>
<td>shot</td>
<td>sled</td>
</tr>
<tr>
<td>arm</td>
<td>beau</td>
</tr>
<tr>
<td>drake</td>
<td>fill</td>
</tr>
<tr>
<td>jaunt</td>
<td>made</td>
</tr>
<tr>
<td>probe</td>
<td>ought</td>
</tr>
<tr>
<td>sign</td>
<td>smash</td>
</tr>
<tr>
<td>beam</td>
<td>change</td>
</tr>
<tr>
<td>dull</td>
<td>fold</td>
</tr>
<tr>
<td>kit</td>
<td>mood</td>
</tr>
<tr>
<td>prod</td>
<td>owe</td>
</tr>
<tr>
<td>snow</td>
<td>smooth</td>
</tr>
<tr>
<td>bliss</td>
<td>climb</td>
</tr>
<tr>
<td>feel</td>
<td>for</td>
</tr>
<tr>
<td>lag</td>
<td>mop</td>
</tr>
<tr>
<td>punk</td>
<td>patch</td>
</tr>
<tr>
<td>sprig</td>
<td>soap</td>
</tr>
<tr>
<td>chunk</td>
<td>corn</td>
</tr>
<tr>
<td>fine</td>
<td>gem</td>
</tr>
<tr>
<td>latch</td>
<td>moth</td>
</tr>
<tr>
<td>purse</td>
<td>pelt</td>
</tr>
<tr>
<td>spy</td>
<td>stead</td>
</tr>
<tr>
<td>clash</td>
<td>curb</td>
</tr>
<tr>
<td>frisk</td>
<td>grape</td>
</tr>
<tr>
<td>loss</td>
<td>muff</td>
</tr>
<tr>
<td>reef</td>
<td>plead</td>
</tr>
<tr>
<td>stiff</td>
<td>taint</td>
</tr>
<tr>
<td>code</td>
<td>deaf</td>
</tr>
<tr>
<td>fudge</td>
<td>grave</td>
</tr>
<tr>
<td>low</td>
<td>mush</td>
</tr>
<tr>
<td>rice</td>
<td>price</td>
</tr>
<tr>
<td>tab</td>
<td>tap</td>
</tr>
<tr>
<td>crutch</td>
<td>dog</td>
</tr>
<tr>
<td>goat</td>
<td>hack</td>
</tr>
<tr>
<td>most</td>
<td>my</td>
</tr>
<tr>
<td>risk</td>
<td>pug</td>
</tr>
<tr>
<td>urge</td>
<td>thin</td>
</tr>
<tr>
<td>cry</td>
<td>elk</td>
</tr>
<tr>
<td>have</td>
<td>hate</td>
</tr>
<tr>
<td>mouth</td>
<td>nag</td>
</tr>
<tr>
<td>sap</td>
<td>scuff</td>
</tr>
<tr>
<td>wave</td>
<td>tip</td>
</tr>
<tr>
<td>dip</td>
<td>elm</td>
</tr>
<tr>
<td>hog</td>
<td>hook</td>
</tr>
<tr>
<td>net</td>
<td>nice</td>
</tr>
<tr>
<td>shop</td>
<td>side</td>
</tr>
<tr>
<td>wood</td>
<td>wean</td>
</tr>
</tbody>
</table>

5.2 Performance Evaluation

5.2.1 Test Materials and Evaluation Procedure

The test materials consist of two lists (lists 11 and 13) from the phonetically-balanced word lists published in [13]. Each list consists of 50 monosyllabic English words. The lists were designed so that they have the same difficulty and phonetic composition, reflect the frequency of occurrence in English speech, and contain commonly used English words. The word lists are shown in Table 5.1. The recording of the test materials was performed in a quiet room with a desktop microphone placed about 10 cm in front of the speaker’s lips. The speaker was a male native English speaker.
5.2 PERFORMANCE EVALUATION

Frame Classification

Vowel, Glide, Diphthong

Nasal

Stop

Set a-priori SNR to the residual noise level

Apply a fourth-order LPC analysis to the time-domain frame, and find \( \omega_c \), the center of frequency region with the highest energy concentration

Determine if the spectrum is high-pass or flat \( (5.3) \)

Rayleigh-MMSE Estimator, \( G_{EM} \)

Vowel, Glide, Diphthong

Nasal

Amplify by +3 dB

Multiply the enhanced STSA by a raised cosine window centered at \( \omega_c \) \( (5.1) \)

Plosive

Amplify burst by +6 dB and aspiration by +3 dB

Fricative

If frame is highpass, multiply the enhanced STSA by the highpass filter shown in Fig. 5.1

Amplify by +3 dB

Inverse STFT and Overlap-Add

\( y(n) \)

\( \hat{x}(n) \)

Figure 5.2: Block diagram of the proposed class-based noise reduction system.
5.2. PERFORMANCE EVALUATION

White noise with SNR of 0 dB was added to the clean speech, and the noisy speech was then processed using two schemes: the Rayleigh-MMSE estimator and the proposed Rayleigh-MMSE estimator combined with the class-based cue enhancement scheme. The smoothing factors, $\alpha_\xi$ and $\alpha_W$, were set to be equal in both schemes; $\alpha_\xi = .99$ and $\alpha_W = .9$. Recall also that we have assumed that the system knows the sound class of the noisy input frame. We have performed manual segmentation and labelling on the clean test materials, and this information was given to the system.

In total, four subjects participated in the test (two subjects in each group). All of them were native English speakers, between 20-30 years of age, and had self-reported normal hearing. The subjects were divided into two groups. One group was presented with the words from List 11, processed using the Rayleigh-MMSE estimator, and the words from List 13, processed using the proposed scheme. The list-algorithm combination was interchanged for the other group. Hence, in total, there are 100 words presented to each listener. The words were presented to the listeners one at a time via a set of headphones.

At the beginning of the test, the subjects were told that they would be presented with a list of 100 common English words, and each word was to be played only once. They were instructed to speak, as well as write, what they heard (their responses were voice-recorded for a later analysis purpose). If a word sounded unclear, they were asked to provide their best guess, but an empty response was allowed.

For each subject-list combination, we counted the total number of correct responses and constructed the consonant confusion matrices. The consonant confusion matrix summarizes the numbers of transmitted and received consonant combinations, i.e., it has the following format:
5.2. PERFORMANCE EVALUATION

The analyzed consonants are /p, b, t, d, k, g, s, f, v, f, θ, dʒ, m, n, η/.

### 5.2.2 Results

The total number of correct word recognitions for each subject is summarized in Table 5.2. The shaded portions indicate samples that were processed with the proposed class-based algorithm. The subjects reported that the words were quite hard to recognize, and this was reflected on the low test scores of all the participants. For both lists, the total numbers of correct word recognitions are higher when the proposed algorithm was used. However, the improvements were not always consistently high as to be able to immediately conclude that the proposed algorithm results in an improved intelligibility.

Table 5.2: Total number of correct word recognition obtained from the intelligibility test

<table>
<thead>
<tr>
<th>Group I</th>
<th>Group II</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>List 11</strong></td>
<td>Subject 1</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td><strong>List 13</strong></td>
<td>9</td>
</tr>
</tbody>
</table>

Group II– List 11: Proposed Scheme, List 13: Rayleigh-MMSE.

More results can be drawn from the interclass confusion matrix analysis, shown in Table 5.3. The numbers inside the bracket indicate the numbers of transmitted
5.2. PERFORMANCE EVALUATION

phonemes in each class, and the entries in the table are the average group scores. An improvement in plosive and fricative identification can be seen from the data, especially for the fricative class, where a large improvement can clearly be observed. However, the identification of affricate and nasal classes deteriorates. Misclassification of affricates for fricatives or plosives increases, and nasal identification is also worse when the proposed algorithm is used.

Table 5.3: The interclass confusion matrices obtained from the intelligibility test (Group Average)

<table>
<thead>
<tr>
<th></th>
<th>Group I</th>
<th>Group II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plos. (41)</td>
<td>18.5</td>
<td>3.5</td>
</tr>
<tr>
<td>Fric. (24)</td>
<td>3</td>
<td>4.5</td>
</tr>
<tr>
<td>Affr. (7)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Nasal (11)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Plos. (36)</td>
<td>26.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Fric. (22)</td>
<td>3.5</td>
<td>13.5</td>
</tr>
<tr>
<td>Affr. (5)</td>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>Nasal (15)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The intraclass confusion matrices for plosive and fricative classes obtained from the intelligibility test are presented in Tables 5.4 and 5.5, respectively. The top-right and bottom-left quarters correspond to the phonemes that were enhanced using the proposed algorithm. The intraclass discrimination of plosive sounds does not seem to be enhanced with the proposed algorithm; the off-diagonal entries do not seem to diminish with the proposed algorithm. For fricatives, the recognition of /s/ was greatly enhanced when the proposed algorithm was used, and in List 13, a perfect recognition of /s/ was achieved. More extensive testing that includes more phoneme occurrences and subjects would be needed to conclude the intraclass recognition performance of
5.3. SUMMARY AND CONCLUSIONS

In this chapter, we have presented the implementation of the proposed class-based noise reduction system. The implementations of the acoustic cue enhancement blocks have been presented in Section 5.1. The summary of this system can be seen in Figure 5.2. We have assumed that the class information is known, and this information was obtained by performing manual labelling on the clean speech sample.

We have also tried to use different estimators for different sound classes, in particular, frames that are tagged as fricatives, burst, or aspiration were processed using the Rayleigh-MMSE estimator, and the others were processed using the Gamma-MAP estimator. The resulting speech sounded unnatural since segments processed with

Table 5.4: The plosive intraclass confusion matrices obtained from the intelligibility test (Group Average)

| Transmitted | Received | Group I | | Group II | | | |
|---|---|---|---|---|---|---|
| | | p | b | t | d | k | g | p | b | t | d | k | g |
| List 11 | | | | | | | | | | | | | |
| p (11) | 4 | 2 | 2 | 2 | 2 | 0.5 | 1 | 1 | 1 | 1.5 |
| b (4) | 1 | 1 | 1 | 1 | 1 | |
| t (5) | 1.5 | 1 | 1 | 1 | 1.5 | 1 |
| d (7) | 1 | 3.5 | |
| k (10) | 1 | 3.5 | 2.5 | 1 | 1 |
| g (4) | 1 | 3.5 | |
| List 13 | | | | | | | | | | | | | |
| p (10) | 2.5 | 2.5 | 2 | 2 | 3.5 | 2 |
| b (3) | 2 | 2 | 1 | 1 | 1 |
| t (8) | 1.5 | 1 | 2 | 1 | 1 |
| d (5) | 3 | 1 | 1 | 2 |
| k (5) | 1 | 3 | 2 | 1 | 1 | 0 | 1.5 |
| g (5) | 3 | 2 | 1 | 1 | 1.5 |
Table 5.5: The fricative intraclass confusion matrices obtained from the intelligibility test (Group Average)

<table>
<thead>
<tr>
<th>Transmitted</th>
<th>Received</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group I</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>List 11</td>
<td></td>
</tr>
<tr>
<td>s (13)</td>
<td>2</td>
</tr>
<tr>
<td>f (4)</td>
<td>1 0.5</td>
</tr>
<tr>
<td>v (2)</td>
<td>0.5</td>
</tr>
<tr>
<td>j (4)</td>
<td>1 0.5</td>
</tr>
<tr>
<td>e (1)</td>
<td>1</td>
</tr>
<tr>
<td>List 13</td>
<td></td>
</tr>
<tr>
<td>s (10)</td>
<td>8.5</td>
</tr>
<tr>
<td>f (6)</td>
<td>2 0</td>
</tr>
<tr>
<td>v (1)</td>
<td>0.5</td>
</tr>
<tr>
<td>j (2)</td>
<td>1</td>
</tr>
<tr>
<td>e (3)</td>
<td>1.5</td>
</tr>
</tbody>
</table>

In Section 5.2, the performance of the proposed class-based noise reduction system was compared with that of the Rayleigh-MMSE estimator in terms of the intelligibility of their output speech. Subjects were presented with noise-reduced samples processed by the Rayleigh-MMSE estimator and the proposed algorithm. The original SNR was 0 dB, and the noise type was white. The consonant confusion matrices were then constructed based on the subjects’ responses (Tables 5.3, 5.4, and 5.5). We concluded that the proposed algorithm was able to improve listeners’ abilities to recognize plosive and fricative classes. However, the intraclass discrimination of plosive sounds did not appear to improve. Furthermore, a more extensive evaluation needs to be done to examine the fricative intraclass recognition performance of the proposed algorithm. The misclassification rate of affricates for plosives or fricatives was higher, and the detection of nasal sounds deteriorated when the proposed algorithm was used.

Hence, we concluded that the proposed class-based processing strategies targeted
at enhancing the plosive and fricative interclass discrimination cues were quite useful. The strategies aimed at enhancing intraclass plosive discrimination, however, did not exhibit a meaningful improvement. The highpass spectrum shaping aimed to improve the recognition of /s/ sound might also be useful, but more extensive testing would need to be performed to conclude so.
Chapter 6

Conclusions

6.1 Summary

The main objective of our work was to incorporate acoustic-phonetic knowledge in the design of a noise reduction algorithm with the aim to recover or enhance acoustic cues which are important for speech intelligibility. To achieve this goal, we considered the following questions: 1) what is the noise reduction algorithm that should be used, 2) what are the acoustic cues that should be enhanced, and 3) how to incorporate this knowledge into the design of the noise reduction algorithm. The first question was explored in Chapter 3 and 4 of this thesis. The second and third questions were explored in Chapter 2, in the form of a literature review and the proposed system overview. Chapter 5 discussed the intelligibility evaluation of the proposed system.

We employed an STFT-domain, Bayesian noise reduction algorithm, similar to that proposed by Ephraim and Malah in [16]. This algorithm assumes a known a-priori distribution for the speech spectral amplitude. The original algorithm assumed
that the STFT components of both speech and noise have complex Gaussian distributions (or equivalently, their STFT magnitude and phase are independent and distributed according to the Rayleigh and Uniform PDFs). However, this was assumed mostly for the tractability of the resulting estimator. In Chapter 3, we evaluated the goodness-of-fits of several parametric PDFs to the empirical speech STSA data. We concluded that for classified speech, Rayleigh and Gamma($v = 5$) PDFs model the speech STSA equally well.

Chapter 4 discussed the noise reduction algorithms. First, a literature review of Bayesian noise reduction algorithms and the Rayleigh-MMSE estimator were presented. We then provided the derivation of the proposed estimators, i.e., the Gamma-MAP and the Gamma-MMSE estimators. These estimators assume that the speech STSA has a Gamma distribution, and the STFT phase is uniformly distributed and independent of the magnitude. The estimators were derived according to the MAP (maximum a-posteriori) and MMSE (minimum mean squared error) criteria. The subjective and objective performances of these estimators were then evaluated. We concluded that, based on the subjective evaluation, the choice of estimator to be used is not so straightforward—listeners’ estimator preferences vary according to environmental conditions and depend on many factors such as the SNR level and noise type. Based on the objective criteria, i.e., the signal distortion and noise reduction factors defined in (4.37) and (4.38), the Wiener, Gamma-MAP, Rayleigh-MMSE, and Gamma-MMSE estimators cause the most to the least amount of noise reduction and signal distortion. Based on this, we proposed to apply the Rayleigh-MMSE estimator to process plosive, fricative, and affricate frames, and apply the Gamma-MAP estimator to process vowel and nasal frames since they have more steady characteristics.
and, intuitively, can afford more signal distortion. However, a simulation of this estimator produced a signal that sounded quite unnatural since the segments processed with the Rayleigh-MMSE estimator contained much more residual noise compared to those processed with the Gamma-MAP estimator. The transition from one estimator to another was reflected in an abrupt sound transition in the output speech.

We also proposed to perform a class-based cue-enhancement processing, in addition to the statistical noise reduction algorithm. We assumed that the system has a knowledge of the sound class of each input frame. In our simulation, this information was obtained by performing manual segmentation and labelling of the test samples. The class-based processing scheme is described as follows.

- For plosives, the following actions were undertaken:
  - perform strong noise suppression to stop segments by setting the a-priori SNR (ξ) parameter in the estimator to the residual noise level,
  - enhance the place of articulation cue in the burst and aspiration segments by finding the region in the spectrum that has the highest energy concentration. This analysis was performed on the noisy input frame. A spectrum shaping was then applied to the enhanced STSA,
  - enhance the burst and aspiration cues by amplifying the enhanced frames by +6 and +3 dB, respectively.

- For fricatives, we aimed to enhance the place of articulation cue by performing an analysis on the noisy input frame to determine if the spectrum has a highpass or flat trend. If it has a highpass trend, a highpass filter is applied to the enhanced STSA. The frication cue was enhanced by amplifying the enhanced
6.2. FUTURE WORK

STSA by +3 dB.

- Affricates were processed as a combination of a burst segment followed by a fricative segment.

- Nasality cue was enhanced by amplifying the enhanced signal by +3 dB.

The implementation of the proposed class-based processing was described in Chapter 5.

The intelligibility of noisy speech processed by the proposed algorithm and the Rayleigh-MMSE estimator was compared in Chapter 5. The test materials are two lists from the phonetically-balanced word lists published in [13]. Based on subjects' responses, the consonant confusion matrices were constructed. From them, we concluded that the proposed algorithm was quite successful in increasing the detection of plosives and fricatives, although intraclass discrimination of plosive sounds did not appear to be enhanced, and more tests need to be done to conclude the intraclass discrimination performance of fricative sounds. The recognition of nasals and affricates, however, deteriorated.

The next section discusses some suggestions on future work.

6.2 Future Work

A potential topic that can be investigated is the use of different kinds of noise reduction algorithms. In particular, we considered to use the suboptimal Wiener filter proposed in [10], where the amount of noise reduction and signal distortion can be controlled by the user. Another suggestion to improve the proposed system is to perform a segmentation-based processing instead of a fixed-rate frame-based processing.
In segmentation-based processing, signal modification is performed to a segment of speech which is approximately stationary. The segments vary in length, and segment boundaries correspond to phonemic or subphonemic boundaries [28]. In this way, the estimation of the gross spectrum trend in the burst/aspiration and fricative segments might be improved. Furthermore, in Chapter 2, we have presented other potential acoustic cues that can also be enhanced, but were not explored in the proposed noise reduction system.
Bibliography


[16] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square
error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics,

models for enhancing noisy speech. *IEEE Transactions on Acoustics, Speech,

[18] Y. Ephraim and H.L. Van Trees. A signal subspace approach for speech enhance-

of noisy speech. *IEEE Transactions on Speech and Audio Processing*, 1:84–89,
January 1993.


[21] V. Hazan and A. Simpson. The effect of cue-enhancement on the intelligibility of
nonsense word and sentence materials presented in noise. *Speech Communication*,

[22] V. Hazan and A. Simpson. The effect of cue-enhancement on consonant intelli-
gibility in noise: speaker and listener effects. *Language and Speech*, 43:273–294,
2000.


