Intelligibility model optimisation approaches for speech pre-enhancement

Maryam Mohamed Al Dabel

Department of Computer Science

The University of Sheffield

A dissertation submitted in partial fulfilment of
the requirements for the degree of

Doctor of Philosophy

Supervisor: Dr. Jon Barker

December 2016

I would like to dedicate this thesis to my parents and my beloved husband for their love, support and encouragement...

Declaration

I hereby declare that this thesis is my own work and effort and that it has not been submitted anywhere for any award. Where other sources of information have been used, they have been acknowledged.

Maryam Mohamed Al Dabel
December 2016

Acknowledgements

In the name of Allah, the Most Gracious and the Most Merciful. Thanks to Allah, who is the source of all the knowledge in this world, for the strengths and guidance in completing this thesis. It is a pleasure to attribute credit to the many people who have contributed directly or indirectly to this work.

First of all, I would like to express my deepest gratitude to my supervisor Dr Jon Barker for his guidance and invaluable advice. His constructive comments and suggestions, throughout the experimental and thesis work, have contributed to the success of this research. His timely and efficient contribution helped me shape this thesis into its final form. Special thanks to my panel members, Professor Guy Brown and Dr Steve Maddock for their useful suggestions. I would also like to take this opportunity to thank Dr Stuart Cunningham and Professor Yannis Stylianou - my viva examiners, for their very helpful comments and suggestions.

My parents for their support, prayers, endless love and care throughout my life. I owe a lot to them, they encouraged and helped me at every walk of my life. I owe a deep sense of gratitude to my beloved husband, Mohammed, for his love, encouragement and constant moral and financial support. He pushed me out through the difficult moments of the study and always motivated me. I would like to extend my deepest gratitude to my children Esknder and Elias, my brother Jassem, my sisters Rabab, Hajer, Zahra, Batool and Fatimah, relatives and friends for their continuous support and prayers.

I thank those who rendered their direct or indirect support during the period of my PhD work. The members of the Speech and Hearing group at Sheffield University have been, and continue to be, a source of knowledge, friendship and humour. My friends, especially those who never let me feel that I am away from my homeland, for their kindness, well wishes and constant prayer. I am also grateful to many individuals at university of Sheffield who provided me with academic and technical support during my research.

Lastly, I gratefully acknowledge the financial support for my research from the King Abdullah Scholarship Program and the government of Saudi Arabia, who gave me this

scholarship to pursue my PhD studies.

Abstract

The goal of improving the intelligibility of broadcast speech is being met by a recent new direction in speech enhancement: near-end intelligibility enhancement. In contrast to the conventional speech enhancement approach that processes the corrupted speech at the receiver-side of the communication chain, the near-end intelligibility enhancement approach pre-processes the clean speech at the transmitter-side, *i.e.,* before it is played into the environmental noise. In this work, we describe an optimisation-based approach to near-end intelligibility enhancement using models of speech intelligibility to improve the intelligibility of speech in noise.

This thesis first presents a survey of speech intelligibility models and how the adverse acoustic conditions affect the intelligibility of speech. The purpose of this survey is to identify models that we can adopt in the design of the pre-enhancement system. Then, we investigate the strategies humans use to increase speech intelligibility in noise. We then relate human strategies to existing algorithms for near-end intelligibility enhancement. A closed-loop feedback approach to near-end intelligibility enhancement is then introduced. In this framework, speech modifications are guided by a model of intelligibility. For the closed-loop system to work, we develop a simple spectral modification strategy that modifies the first few coefficients of an auditory cepstral representation such as to maximise an intelligibility measure. We experiment with two contrasting measures of objective intelligibility. The first, as a baseline, is an audibility measure named 'glimpse proportion' that is computed as the proportion of the spectro-temporal representation of the speech signal that is free from masking. We then propose a discriminative intelligibility model, building on the principles of missing data speech recognition, to model the likelihood of specific phonetic confusions that may occur when speech is presented in noise. The discriminative intelligibility measure is computed using a statistical model of speech from the speaker that is to be enhanced.

Interim results showed that, unlike the glimpse proportion based system, the discriminative based system did not improve intelligibility. We investigated the reason behind that and we found that the discriminative based system was not able to target

the phonetic confusion with the fixed spectral shaping. To address that, we introduce a time-varying spectral modification. We also propose to perform the optimisation on a segment-by-segment basis which enables a robust solution against the fluctuating noise. We further combine our system with a noise-independent enhancement technique, *i.e.,* dynamic range compression. We found significant improvement in non-stationary noise condition, but no significant differences to the state-of-the art system (spectral shaping and dynamic range compression) where found in stationary noise condition.

Contents

List of Figures

List of Tables

Nomenclature

Roman Symbols

$\mathcal{C}$	Constant
c_n	Cepstral coefficients
C_t	Constant for each time-frame t
$\mathbf{c}_t$	Cepstral coefficients within each time-frame t
$\mathbf{c}_{t_i}^{j}$	Cepstral coefficients for a time-frame t_i of segment j
$C_{t_i}^{j}$	Constant for a time-frame t_i of segment j
$\mathcal{D}$	Discriminative microscopic intelligibility
$e(t)$	Temporal envelope of speech signal before applying DRC
$\bar{e}(t)$	Temporal envelope of speech signal after applying DRC
F	Total number of frequency channels
f	Frequency-band
f_0	Fundamental frequency
f_b	3-dB bandwidth
f_c	Centre frequency
f_f	Centre frequency of the band f in Hz
f_s	Sampling frequency
G_1	First-order complex bandpass frequency response

K_4 Fourth-order complex bandpass frequency response

g_f Band-dependent weights

$\hat{\mathbf{c}}$ Optimal cepstral coefficients

$\mathcal{H}(.)$ Heaviside step function

$\hat{x}$ Enhanced speech signal

$\hat{X}_c^j(t_i, f)$ Modified and re-normalised speech spectrum of segment j

I^j Total number of time-frames within a defined segment j

J Total number of segments of speech spectrum segments

k Mixture component

M Total number of mixture component

$M(t, f)$ Glimpse mask

$\mathcal{S}_c$ Weighting matrix of the same size $(T \times F)$

$\mathcal{S}_c^j$ Weighting matrix of the same size as segment j $(I^j \times F)$

n Noise signal

$N(t, f)$ Noise spectrum

$N'(t, f)$ Scaled noise spectrum

$N^j(t_i, f)$ Noise spectrum of segment j

$n(t, f)$ Noise energy within the element of time frame t and frequency channel f

$\mathbf{x}$ Acoustic feature vector

$Q^{(1)}$ Correct state sequence

$Q_{t_i}^{(1)}$ Correct state sequence within each time-frame, t_i, for segment j

$Q^{(2)}$ Best incorrect state sequence

$Q_{t_i}^{(2)}$ Best incorrect state sequence within each time-frame, t_i, for segment j

R_c Root-mean-square after applying the optimal spectral shaping to speech spectrum

$x'_f(t)$ Real part of the impulse response (fine structure)

X^r Reliable features

$S_c(t, f)$ Weight for each time-frame t

$S_c^j(t_i^j, f)$ Weight for a time-frame t_i of segment j

$S_c(f)$ A band-dependent scaling

T Total number of time frames

t Time-frame

$\tilde{b}_f$ Frequency band-dependent complex factor

τ_f Either the point in time of the envelope maximum of the respective impulse response or the desired group delay

t_i^j Time-frame of segment j

$\tilde{x}_f(t)$ Complex filterbank output signals

X^u Unreliable or missing features

x Speech signal

$X(t, f)$ Speech spectrum

$\hat{X}_c(t, f)$ Modified and re-normalised speech spectrum

$\tilde{x'}_f(t)$ Complex filterbank output signals after being multiplied by $\tilde{b}_f$

$X^j(t_i, f)$ Clean speech spectrum of segment j

$x'''(t)$ Resynthesised speech signal

$x(t, f)$ Speech energy within the element of time frame t and frequency channel f

$x''_f(t)$ Real part after being delayed by Δt_f

y Noisy speech signal

$Y_c(t, f)$ Noisy enhanced spectrum

Greek Symbols

β Oscillation frequency of the filter (phase)

δ Optimised parameters

γ Filter order

λ Acoustic speech model

λ_c Acoustic speech model after applying spectral shaping

μ Means vector

μ_c Means vector after applying spectral shaping

π $\simeq 3.14\ldots$

Σ Covariance matrices

σ_e Standard Error

τ Time constant

θ A predefined threshold

φ Bandwidth parameter

Acronyms / Abbreviations

3DDS Three-Dimensional Deep Search

AI Articulation Index

ANOVA Analysis of variance

ASA Auditory Scene Analysis

ASR Automatic Speech Recogniser

BMN Babble Modulated Noise

CELP Code Excited Linear Prediction

CM Confusion Matrices

CP Confusion Pattern

CSII Coherence Speech Intelligibility Index

CSTI Normalised Covariance based Speech Transmission Index

CVC Consonant-Vowel-Consonant

CVs Consonants-Vowels

DAU Dau auditory model

DC Degree of Change

DIS Discriminative microscopic intelligibility model

DIS-OPT DIS-optimised speech using spectral shaping modification

DTW Dynamic-Time-Warp

EM Energetic masking

ERB Equivalent Rectangular Bandwidth

ESII Extended Speech Intelligibility Index

F1 First formant frequencies

F2 Second formant frequencies

FFT Fast Fourier Transform

GF Gammatone Filterbank

GMM Gaussian mixture models

GP Glimpse Proportion

GP-OPT GP-optimised speech

HMM Hidden Markov Models

IIR Infinite impulse response

IM Informational Masking

IOEC Predefined input/output envelope characteristic

LPC Linear Predictive Coding

MSC Magnitude Squared Coherence

MTF Modulation Transfer Function

NCM Normalised Covariance Metric

NM Nelder-Mead Simplex Optimisation

NSEC Normalised Subband Envelope Correlation

DRC Dynamic range compression

ORG-DRC ORG modified with dynamic range compression

ORG Original unmodified speech

SD Speaker Dependent acoustic model

sEPSM Speech-based Envelope Power Spectrum Model

SII Speech Intelligibility Index

SI Speaker Independent acoustic model

SNR Signal-to-Noise Ratio

SPL Sound Pressure Level

SRT Speech Reception Threshold

SSDRC Spectral shaping combined with DRC

SSN Speech-Shaped Noise

SS Spectral Shaping based modified speech

STEP Spectro-Temporal Excitation Pattern

STI Speech Transmission Index

STMI Spectro-Temporal Modulation Index

STOI Short-Term Objective Intelligibility

S-T Spectro-Temporal elements

TTS Text-To-Speech systems

TVDIS-DRC TVDIS-optimised speech combined with DRC

TVDIS DIS-optimised speech using time-varying spectral modification

TVGP-DRC TVGP-optimised speech combined with DRC

TVGP GP-optimised speech using time-varying spectral modification

VCV Vowel-Consonant-Vowel

Chapter 1

Introduction

Imagine walking through a busy train station while being late for a train and trying to listen for an announcement *e.g.,* "The train on Platform 3 is the 9am to Sheffield". The most important part to understand is when and from which platform your train is going to leave. This message is typically transmitted from a digital communication system, a so-called public address system, which may have been designed to be easily intelligible in an ideal listening scenario (*i.e.,* an unrealistic listening situation). However, in a real listening situation, intelligibility is often degraded due to the presence of noise in the environment of the listener to whom the message is being delivered. Generally speaking, there is an opportunity, somewhere in this communication system, to access the clean speech signal before it is corrupted by the environmental noise. There is therefore an opportunity to *pre-enhance* the clean speech signal *before* it is transmitted. If the pre-enhancement is appropriately designed, it may be possible to preserve the intelligibility of the signal, *i.e.,* making it immune to the effects of the noise.

In this thesis, we aim to design and evaluate a *better* pre-enhancement system to improve speech intelligibility in noise. To do so, we will use models of speech intelligibility.

1.1 What is Speech Intelligibility?

The perception of a speech signal is often measured in terms of its intelligibility and quality. Intelligibility is the degree to which the message can be *understood*. For a message to be understood, it is usually required that the sub-units (*i.e.,* phones, syllables, words or sentences) can be correctly identified. Note, intelligibility is not the same as speech quality which is a subjective measure that summarises on individual preferences of human listeners (Kondo, 2012). There is not a clear relationship between the two

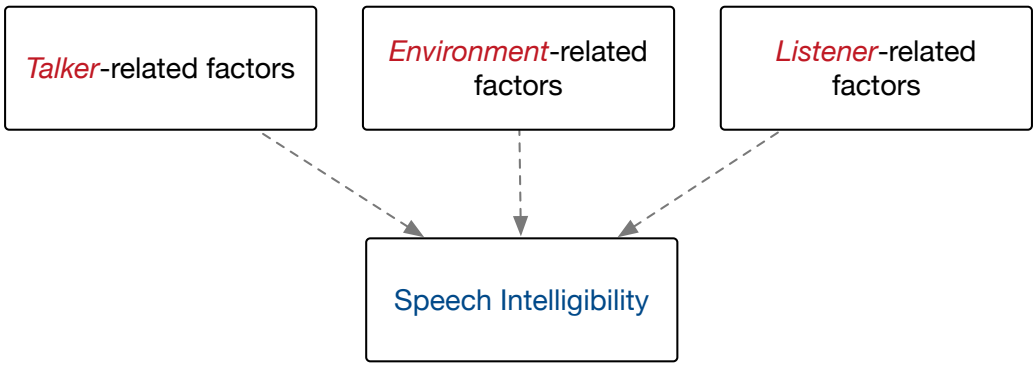

Figure 1.1: A schematic diagram of factors influencing speech intelligibility including: talker-related, environment-related, and listener-related factors.

measures. So, one could improve the intelligibility of speech at the expense of sacrificing the quality. Improving the quality, on the other hand, does not necessarily increase speech intelligibility (Ephraim and Cohen, 2005). In this thesis, we concentrate on improving speech intelligibility.

The relative importance of intelligibility verses quality depends on the listening situations. For example, listening to an announcement in the train station would be an example scenario representing the importance of intelligibility over the quality. A common situation of the importance of quality over the intelligibility would be listening to poetry or music.

There are a number of aspects of the speech communication process that should be considered for improving the overall intelligibility of speech. From a signal processing perspective, the communication process can be modelled as three stages: (i) speech production by the talker (*i.e.,* transmitter), (ii) transmission of the signal over a channel, and (iii) speech perception by the listener (*i.e.,* receiver). The process can be summarised as follows: a message is generated and encoded in language in the talker's brain. The message in then modulated into an air pressure wave to yield an acoustic speech signal. The speech is then transmitted into the transmission channel which might be an acoustic environment or the speech might be transmitted into an electrical signal (*e.g.,* telephone communication channel). The speech signal is then possibly degraded by the transmission channel depending on factors such as whether the environment is quiet, whether it contains additional noise sources, and whether the environment is reverberant. The signal is then received by the listener's ear in which it is demodulated into a message

that is decoded by the brain.

Sources of adversity may arise during production, reception, or during the transmission channel that may lead to reduction in intelligibility. As displayed in Figure 1.1, these factors can be summarised as follows: talker-related, environment-related, or/and listener-related.

In terms of talker-related factors, the talker usually deliberately or unconsciously adapts his/her speaking style in order to maintain intelligibility in face of changes in the communication environment, *i.e.,* the context. The context might be noisy environment, or when speaking with particular group of listeners, *e.g.,* a listener with hearing impairment or a non-native listener (*i.e.,* cognitive influences of speaking environment). This process can break down in certain situations, *e.g.,* if the listener and talker are not in the same environment. Other talker-related difficulties may include: language differences, accents, speed, speech styles, degree of articulation, talker variability, speech dis-order, to name a few.

Moving to environment-related factors, there are several factors in the environments that affect the intelligibility of speech (*e.g.,* (Assmann and Summerfield, 2004)). Example factors may includes differences in the number of sources, the location of the sources, distance between the sources, and the movement of the sources. Furthermore, the room geometry, the reverberation of room, and the distance from the talker to the listener may also influence speech intelligibility.

Intelligibility is also a function of listener. The ability of humans listeners to interpret the speech depends on auditory and cognitive processes. It will be influenced by the listeners' inability to segregate audible parts of the target signal from masker. For this reason, this ability differs among listeners according to factors such as hearing impairment including age-related hearing loss. Additional factors may include language, accent, and talker familiarity.

In summary, there are lots of factors associated with each stages of the system. This means that for one to build an algorithms to model speech intelligibility, it is necessary to consider many related factors, *i.e.,* having a model of the talker, a model of the environment, and a model of the listener. The development of such algorithms, however, has remained to be error-proofing, this was due to the fact that algorithms were sought that would work for all the factors which is an extremely challenging task. Despite the complexity in modelling all the factors, efforts to estimate intelligibility objectively without humans subjects do exist and these will be reviewed in Chapter 2 and used in this thesis.

1.2 Speech Near-end Intelligibility Enhancement

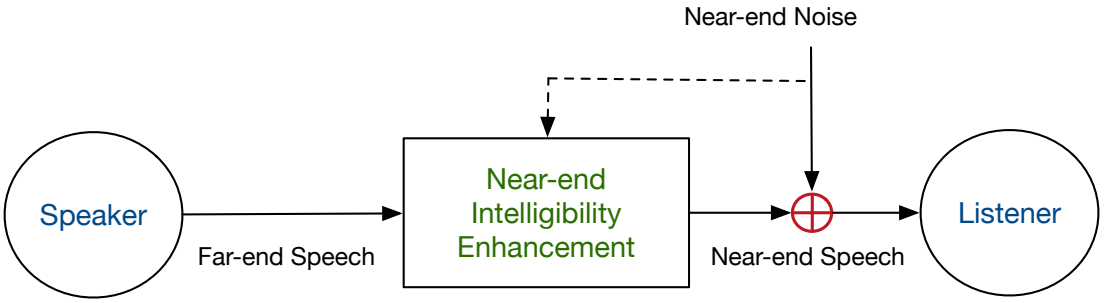

Figure 1.2: A schematic diagram of near-end intelligibility enhancement system.

In the conventional noise suppression techniques, the distorted speech is processed *at the receiver-side* in order to estimate and reconstruct the clean speech signal. On the contrary, the pre-enhancement, also known as near-end intelligibility enhancement, covers a set of techniques which pre-process the clean far-end speech *at the transmitter-side*, *i.e.,* before it is played into the communication channel (*i.e.,* the near-end noise), in order to improve its intelligibility for the near-end listener. A schematic of the near-end intelligibility enhancement is shown in Figure 1.2.

The ultimate goal of the near-end intelligibility enhancement is to increase the signal's intelligibility, while hopefully maintaining its quality. Generally speaking, a simple speech modification solution to make speech more intelligible in noisy environment would be to amplify the power of speech signal before transmission to the listening environment, and therefore increasing the signal-to-noise ratio (SNR) (Brouckxon et al., 2008). Turning the volume up is clearly an effective solution in some situations but undesirable due to a number of sides effects including: (i) it requires more energy; (ii) it may lead to hearing damage (Sabin and Schoenike, 1998); (iii) it may cause distortion of the speech signal; and finally (iv) it may be socially unacceptable if the communication channel is a shared-channel (*e.g.,* speech being broadcast acoustically). This thesis will be devoted to increasing speech intelligibility in noise under an equal energy constraint, *i.e.,* the energy of speech is constrained to remain the same before and after modification.

The operation of the near-end intelligibility enhancement and its design is inspired by the fact that talkers adjust their speech to the available listening context, *e.g.,* background noise. Therefore, the vast majority of near-end intelligibility enhancement techniques attempt to model context-aware speech production by using one of two systematic approaches: open-loop and closed-loop *feedback* systems, as depicted in Figure 1.3.

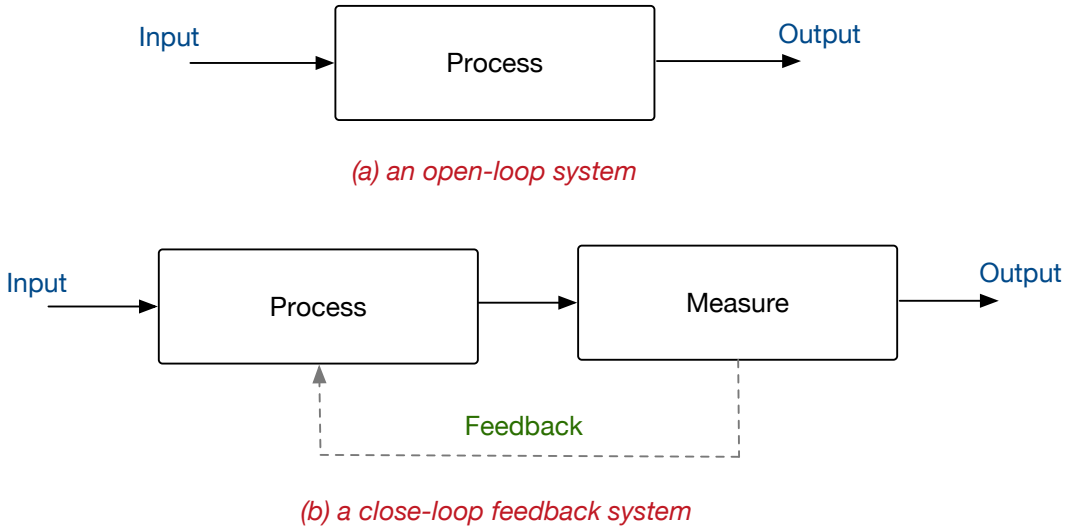

Figure 1.3: A schematic diagram of approaches taken for speech near-end intelligibility enhancement system: (a) an open-loop system, and (b) a closed-loop feedback system.

The open-loop system is often represented as multiple cascaded subsystems in series or just a single system with an input and output signal. It demonstrates a linear path from the input contextual information to the output speech signal with no feedback loop. Whereas, in a closed-loop framework, feedback is embedded to ensure the design and tuning strategy of the modification are appropriate. For that, one requires to: (i) determine the desired control loop performance, this could for instance be carried out by optimising intelligibility score; (ii) determine a model to be controlled, this may include, for example, an objective intelligibility measure, or an automatic speech recogniser (ASR) system; and (iii) choose a suitable modification design method in order to make it possible to achieve the desired performance for the chosen model under an energy-preservation constraint. Additionally, at least two inputs are required for this system including information about the spoken message and information about the context.

In the closed-loop approach, the parameters of the modification are generally updated based on the current estimate by the model. Thus, it provides the following advantages over the open-loop approach including: (i) the opportunity of getting better parameter estimates for modification design; (ii) the opportunity of re-initialisation of the modification parameters after each updating without opening the loop; and (iii) the opportunity of using a more sophisticated modification design procedure although it may requires a large amount of computation. For these reasons, we adopt the closed-loop feedback

framework for developing the new near-end intelligibility enhancement system presented in this thesis.

1.3 Thesis aim

The main aim of this thesis is to answer the question: How can we modify speech by exploiting a *priori* knowledge about a talker and the noise environment while using a closed-loop optimisation approach? To achieve this aim, we concentrate on a combination of four research ideas: (i) exploiting advanced models of intelligibility; (ii) personalisation of intelligibility models to adapt the intelligibility benefit to the near-end listeners; (iii) exploiting knowledge of the near-end noise source; (iv) designing flexible speech modification strategies. The combination of these four research ideas provides the possibility for optimising the intelligibility of speech. Although objective intelligibility models that provide a numeric estimate of intelligibility in a given listening condition can be helpful, particularly during the development stage of near-end intelligibility enhancement algorithms, it is crucial that the final validation of intelligibility improvement use listening tests with human subjects. Thus, we aim to assess the eventual output of the closed-loop based framework using such tests.

1.4 Thesis Structure

The remainder of this thesis is structured as follows:

In Chapter 2, we present a survey of speech intelligibility models and how the adverse acoustic conditions affect the intelligibility of speech. The purpose of this survey is to identify models that we can adopt in the design of the enhancement system. Then, in Chapter 3 we investigate the strategies humans use to increase speech intelligibility in noise. We then relate human strategies to existing algorithms for near-end intelligibility enhancement, categorised into those designed for noise-only, reverberation-only, and finally for a combination of additive noise and reverberation.

Having surveyed models of intelligibility, and possible modification strategies, we introduce a closed-loop feedback approach to near-end intelligibility enhancement. Chapter 4 provides a detailed account of a framework that modifies speech using a model of intelligibility inspired by the observation of humans' natural speech enhancement strategies. We refer to this framework as an *analysis-resynthesis* framework. The analysis-resynthesis framework has two distinct tasks. The analysis stage extracts the acoustic

parameters of the input speech. The resynthesis stage is responsible for waveform regeneration where acoustic parameters extracted from the analysis are then used to construct a speech waveform. The chapter presents a simple example that confirms the applicability of the implementation.

In Chapter 5, we propose a new near-end intelligibility enhancement system to improve the intelligibility of speech in noise. We develop a simple spectral modification strategy that modifies the first few coefficients of an auditory cepstral representation such as to maximise an intelligibility measure known as the glimpse proportion (GP). We describe the GP and the spectral modification in more depth. We then integrate this measure into an existing optimisation method for spectral modification. Acoustic analyses, objective experiments, and listening tests are presented.

The GP is a *blind* measure that does not take into account any speech or speaker-specific knowledge, in particular, it does not model the likelihood of specific phonetic confusions that may occur when speech is presented in noise. To address this weakness, in Chapter 6 we propose a discriminative intelligibility model, building on the principles of missing data speech recognition (*e.g.,* Cooke et al. (2001)). We also describe how this model and the spectral modification fit into the framework with the optimisation being performed at an utterance-level. Objective evaluation as well as listening experiments with speech-shaped noise at several SNRs are presented.

To improve the performance of the developed near-end intelligibly enhancement using the spectral modification, we propose in Chapter 7 a *time-varying* spectral modification. We also prepose to perform the optimisation on a segment-by-segment basis which enables a solution for fluctuating noises like babble-modulated noise. Then, we refine the discriminative intelligibility model to also work on a phoneme-level. We also investigate whether it is possible to further improve results by combining the GP-based method with dynamic range compression, a strategy that reallocates energy across different time segments of speech to maximise intensity levels and has been shown to be successful in previous near-end intelligibility enhancement systems. Objective evaluation with acoustic analysis as well as final listening experiments with speech-shaped and babble-modulated noises at several SNRs are presented.

Finally, in Chapter 8, we summarise the main contributions of this thesis and consider the limitations further with a view to possible directions for future work.

Chapter 2

Modelling Speech Intelligibility

2.1 Introduction

The intelligibility of a spoken message, transmitted via speech communication systems, is often decreased at various stages before, during and after transmission. This covers degradations of speech production by the speaker, transmission channel (*e.g.,* channel distortion or masking), and the auditory system of the listener. Thus, to predict speech intelligibility all stages of transmission process (*i.e.,* the speaker, the transmission channel, and listener) should be modelled and that is a challenging task.

The motivation for reviewing speech intelligibility models in this chapter is twofold. Firstly, the review will lead to a better understanding of the mechanisms behind intelligibility models. This understanding will help us selecting a reliable candidate that can fit within a closed-loop near-end intelligibility enhancement system. Secondly, accurate and reliable intelligibility models are of great interest for evaluation of the near-end intelligibility enhancement system. This may replace costly and time-consuming listening tests, particularly, in early stages of the system development process. Therefore, there are two main roles for the intelligibility models used this thesis which are: enhancement process and evaluation process, and thus choosing the appropriate model to use in each process is important.

Attempts to model speech intelligibility have been proposed in the literature. These models generally fall into one of the two main classes depending on the level of details that they attempt to predict: (i) 'macroscopic' modelling aiming at predicting overall speech intelligibility in masking and reverberation; and (ii) microscopic modelling which attempt to predict specific listeners' responses to individual tokens. Each approach has its advantages and disadvantages depending on the application at hand, *i.e.,* whether the

requirement is to provide a good match to average listener recognition rates or matching listener responses at the level of individual tokens (Cooke, 2009).

The remainder of this chapter is outlined as follows: Section 2.2 provides an overview of human perceptual strategies to compensate for speech in additive noise as well as the effects of additive noise on speech intelligibility. Then, we survey the literature in terms of models of speech intelligibility. In particular, Section 2.3 and 2.4 describe the macroscopic intelligibility models and microscopic intelligibility models, respectively. The chapter ends with a summary in Section 2.5.

2.2 The Perception of Speech in Noise

The main purpose of this section is to first identify perceptual strategies that human listeners adopt in the presence of additive noise in order for us to better understand the effects of masking on speech intelligibility, and second to describe these effects. Section 2.2.1 describe the human listening strategies and Section 2.2.2 gives an overview of the differential effects of energetic and informational masking on speech intelligibility. Definitions of energetic and informational masking will be given then.

2.2.1 Human Listening Strategies in Noise

In typical listening conditions, sounds may reach the listener's ear as a mixture of different acoustic sources. Now, to identify individual sound patterns the incoming auditory information need to be organised first, and then the right subset assigned to individual sounds, in order to form an accurate representation for each. This mechanism is named 'auditory scene analysis' (ASA) by Bregman (1990) in which the input acoustics grouped and segregated into separate mental representations, called auditory streams. Bregman (1990) makes a distinction between two types of auditory streaming which are simultaneous grouping and sequential grouping. The simultaneous grouping refers to the process of grouping units occurring all at a certain time but in different spectral bands. In contrast, the sequential grouping is the process of grouping sound units that occur sequentially in time but possibly in the same spectral band.

In addition to identifying the target source, human listeners adopt a strategy that plays a role in source separation named glimpsing. Cooke (2003) defines the glimpsing phenomenon as the ability to extract spectro-temporal elements in which the degraded speech signal is less masked and as a result less distorted. In fact, human listeners utilise

Table 2.1: Summary of potential masking effects for listeners including native and non-native listeners as reported in Cooke et al. (2008).

Energetic masking (EM)	*Informational Masking (IM)*
i) partial information	i) misallocation of audible masker components to target ii) competing attention of masker iii) higher cognitive load iv) interference from 'known language' masker

the local high signal-to-noise ratio (SNR) elements of the noisy signal and obtain useful information, *i.e.,* glimpse, accordingly. For this reason, stationary noises are stronger maskers compared to competing speakers, since the later present more glimpses to the listener (Festen and Plomp, 1990).

2.2.2 Energetic masking and informational masking

When noise interferes with a speech signal (*i.e.,* target) it can provide two types of masking - 'energetic' and 'informational' - either of which can lead to a reduction in intelligibility. Energetic masking (EM) refers to masking which occurs in the periphery of the auditory system when the speech energy in some spectro-temporal region is rendered inaudible owing to the high noise energy. Informational masking (IM) refers to target and masker competition that occurs in more central portions of the auditory system (Durlach et al., 2003). In fact, it is a 'catch-all' term that covers any reduction in intelligibility once energetic masking in the auditory periphery has been accounted for (Cooke et al., 2008; Durlach, 2006).

As the energetic masking increases the speech intelligibility decreases. Energetic masking results in loss of speech observation in spectro-temporal regions which may or may not include important speech features that helps in the discriminability between speech classes.

Informational masking has multiple potential aspects which are summarised in Table 2.1 as reported in Cooke et al. (2008). The first is misallocation of the target source (*i.e.,* which audible components belong to the target source) referring to two situations when: (i) the human listener uses audible elements from the masker leading to misidentification of the target, or (ii) the human listener assigns target elements to the masker leading to erroneous identification, too. Studying the effect of informational masking has been often conducted using speech-like maskers (*i.e.,* contained speech ma-

terial) (Brungart, 2001; Freyman et al., 2004). For example, Brungart (2001) stated that in the presence of a single competing talker all words belonging to the masker are usually reported as part of the target. Simpson and Cooke (2005) also found a substantial effect of informational masking on speech intelligibility when speech presented in **N**-talker babble over a wide range of values for **N**. Therefore, speech-like maskers are more likely to yield this type of masking through misallocation. Cooke et al. (2008) indicated that misallocation could apply to speech sub-units of any size and also speech sub-units smaller than words or phonemes. Misallocation may also lead to report a sound or word which is not parts of either the target or the speech-like masker (*e.g.,* the aspiration comes after a plosive could be perceived as the voiceless glottal fricative /h/) (Cooke et al., 2008).

A further aspect associated with informational masking is the higher cognitive load often occurs when processing a signal that contains multiple components (Cooke et al., 2008). Assuming that both target and masker might have important components, it makes more sense that processing resources are equally assigned to both target and masker. This often occurs in the presence of a competing speech masker and results in the failure to attend to the target. Darwin and Hukin (2000) investigated how differences in properties such as fundamental frequency f_0, vocal tract length, and spatial cues affect detecting which of two competing sentences is more likely attended to. Cooke et al. (2008) also stated that a higher cognitive load is more likely to cause difficulties in tracking the target source specially if attention is resulted from limited resources (*e.g.,* Kahneman (1973)).

An additional effect of informational masking according to Cooke et al. (2008) may arise from the language of the masking talker and whether it is known to listeners. Several recent studies investigating the effect of the language of the masker on the intelligibility of the target sentence (*e.g.,* (García Lecumberri and Cooke, 2006; Rhebergen et al., 2005)). Rhebergen et al. (2005) reported a significant reduction in speech reception thresholds for Dutch sentences presented in competing Dutch speech compared to when presented in competing Swedish speech. In García Lecumberri and Cooke (2006), a consonant in a vowel context identification task were used and they demonstrated that monolingual English listeners were better at identifying the consonant when the language of a competing talker was Spanish. However, Spanish listeners with English as their second language performed equally in the presence of maskers in both languages (*i.e.,* English or Spanish).

Table 2.2: List of macroscopic models of speech intelligibility classified into SNR-based and correlation-based.

SNR-based measures	*Abbreviation*
Articulation Index (French and Steinberg, 1947)	AI
Speech Intelligibility Index (ANSI, 1997)	SII
Speech Transmission Index (Steeneken and Houtgast, 1980)	STI
Spectro-Temporal Modulation Index (Elhilali et al., 2003)	STMI
Glimpses Proportion (Cooke, 2006)	GP
Extended SII (Rhebergen and Versfeld, 2005)	ESII
Speech-based Envelope Power Spectrum Model (Jørgensen and Dau, 2011)	sEPSM

Correlation-based measures	*Abbreviation*
Coherence SII (Kates and Arehart, 2005)	CSII
Normalised Covariance based STI (Goldsworthy and Greenberg, 2004)	CSTI
Normalised Subband Envelope Correlation (Boldt and Ellis, 2009)	NSEC
Dau auditory model (Christiansen et al., 2010)	DAU
Short-Term Objective Intelligibility (Taal et al., 2011)	STOI

2.3 Macroscopic Intelligibility Models

Macroscopic intelligibility models often attempt to predict human intelligibility judgements using long-term statistics of entire speech signals. These models usually comprise a two-stage analysis namely: feature extraction and back-end processing. In particular, features are extracted from the input speech signals to generate an 'internal representation'. The resulting internal representation is then analysed in an independent back-end processing to form a decision metric regarding the signal's intelligibility. The input signals can possibly be: (i) the clean speech, x, and the masker, n, alone, or (ii) the clean speech, x, and the noisy speech, y. Based on that, we classify the macroscopic models into two main classes namely: SNR-based and correlation-based intelligibility models, (a list of these models is shown in Table 2.2). The SNR-based predictions measures assume that speech and masker signals are available in isolation and often use a metric based on the amount of audibility. Models falling under this category are reviewed in Section 2.3.1. The correlation-based predictions measures, on the other hand, consider the similarity between a reference signal (usually clean speech) and a test signal as their decision metric. The latter class is presented in Section 2.3.2.

Table 2.3: List of the SNR-based macroscopic intelligibility models along with their decision matrices using either signal-to-noise ratio (SNR) or modulation transfer function (MTF) in which the processing in the spectral (f), temporal (t) or spectro-temporal (t, f) domains.

study	*measure*	*decision metric*
French and Steinberg (1947)	AI	SNR(f)
ANSI (1997)	SII	SNR(f)
Rhebergen and Versfeld (2005)	ESII	SNR(t, f)
Cooke (2006)	GP	SNR(t, f)
Steeneken and Houtgast (1980)	STI	MTF(t)
Elhilali et al. (2003)	STMI	MTF(t, f)
Jørgensen and Dau (2011)	sEPSM	SNR_{env}

2.3.1 SNR Based Modelling

The primary idea of the SNR-based models is to quantify the amount of distortion caused by the communication channel using a weighting function as its decision metric based on speech and masker signals. Table 2.3 outlines the SNR-based macroscopic intelligibility models along with their decision matrices in which the processing in the spectral, temporal or spectro-temporal domains.

One of the early attempts to define a macroscopic measure was at AT&T Bell Labs around 1920 and finally published by French and Steinberg (1947). The measure was later refined by Kryter (1962) to make it better accessible through proposing a calculation scheme, which is now known as the articulation index (AI). The basic approach of AI is to quantify the degree of energetic masking by estimating the SNR within several frequency bands. The SNRs are then limited to a certain pre-specified SNR range, normalised and subjected to auditory masking effects, and finally coupled by calculating a weighted sum. The AI is scored between zero and unity, such that an AI less than 0.3 indicates poor conditions for listening to speech, while AI of above 0.7 is considered to be excellent.

The AI model was further extended to the speech intelligibility index (SII) and was standardised under S3.5-1997 ANSI (ANSI, 1997). The SII is computed from the speech spectrum, the noise spectrum, and the listener's hearing threshold. The speech and noise signals are filtered into frequency bands, and within each band the factor audibility is derived from the SNR. Bands are then weighted by a weighting function known as the band-importance function that indicates the relative contribution of each frequency band to total speech intelligibility. This is because not all frequency bands contain an equal

amount of information and thus not equally important for speech intelligibility. The outcome of the SII is derived by a weighted sum of the band-importance function across the different frequency bands. The resulting SII is a number between zero and unity. An SII of zero indicates a very poor condition for listening to speech (*i.e.,* no speech information is available), an SII of unity indicates a superb conditions for listening to speech (*i.e.,* all speech information is available).

The SII, however, can only represent intelligibility in cases where background noise is stationary (*e.g.,* Chabot-Leclerc et al. (2014)). To account for non-stationary noises, Rhebergen and Versfeld (2005) proposed an extended version of SII known as extended SII (ESII). The idea behind the ESII was to define an analysis window whose size depends on the temporal resolution of the human ear. Within each time-frame, the conventional SII is determined. The per-frame SII values are then averaged resulting in a final intelligibility prediction.

Cooke (2006) used an approach similar to Rhebergen and Versfeld (2005). Cooke (2006) defined a so-called Glimpse Proportion (GP) measure based on the glimpsing phenomenon (which is a human perceptual strategy that plays a role in source separation). The glimpsing phenomenon is the ability to extract spectro-temporal elements (*i.e.,* glimpses) in which the degraded speech signal is less masked and hence less distorted (Cooke, 2003). The GP measure computed as the percentage of spectro-temporal elements that have a local dB SNR higher than a pre-defined threshold. More information about the computation of this measure will be given in Chapter 5.

The AI and SII were designed for simple linear degradations (*e.g.,* additive noise). To overcome the limitation, Houtgast et al. (1980) proposed the speech transmission index (STI) to predict intelligibility of reverberated speech and non-linear distortions. Conceptually, the STI was derived based on a metric named the modulation transfer function (MTF). According to Houtgast and Steeneken (1973, 1985), the MTF describes the transmission of sound in a room . Its underlying principle is that the signal (which reachs a listener's ear through the room) is not the exact copy of the original source signal instead it is a *blurred* copy. Thus, in the MTF, the room is represented by a linear time-invariant system. Now, in order to quantify the amount of distortion introduced to a signal, the room is tested with a sine in-sine out model, applied in the modulation domain. In particular, a series of test signals (often noise bands across seven octaves of the frequency range between 125 Hz and 8 kHz) are modulated by sine waves covering a range of modulation frequencies from 0.63 to 12.5 Hz in 1/3-octave intervals. Then, the effect of transmission through the room on any reduction in magnitude or modulation

depth in the envelope is reported (Houtgast and Steeneken, 1973, 1985). The STI takes into account the audibility of spectral regions and evaluates reduction in temporal modulation of the speech envelope using the MTF.

However, the STI fails to predict speech intelligibility in nonlinear processing conditions, *e.g.,* envelope compression, phase shifts or spectral subtraction (Chabot-Leclerc et al., 2014). To alleviate this shortcoming in the STI, Elhilali et al. (2003) and Jørgensen and Dau (2011) attempted to defined new models. Elhilali et al. (2003) proposed a spectro-temporal modulation index (STMI). In the STMI, a two-dimensional modulation processing was introduced to account for modulation across the frequency domain as well as the temporal modulations (considered in the STI). Inspired by applying the concept of SNR in the modulation domain in Dubbelboer and Houtgast (2008) and Ewert and Dau (2000), Jørgensen and Dau (2011) introduced a speech-based envelope power spectrum model (sEPSM) as an alternative modification to the STI. The sEPSM measured the SNR in the envelope power domain (SNR_{env}) at the output of a modulation filterbank. The underlaying principle of the SNR_{env} metric was that an estimate of the speech-alone modulations can be derived from the modulation spectrum of the noisy speech and the noise alone (instead of the reduction in modulation due to distortion). Results showed that in contrast to the SII and STI, the sEPSM considered the intelligibility of speech degraded with speech-shaped masker as well as for the effects of noise reduction using spectral subtraction algorithm and the effects of reverberation as reported in (Dubbelboer and Houtgast, 2007; Hilkhuysen et al., 2014).

The SNR-based models were considered to be useful for several types of distortions including additive noise, reverberation, filtering, and clipping. However, they fail in situations where speech signals are degraded by non-stationary noise sources and processed by time-varying and non-linear filtering systems such as those often embedded in single-channel speech enhancement systems (Taal et al., 2010).

2.3.2 Correlation Based Modelling

Correlation-based intelligibility models require both the clean and the noisy inputs signals to derive a weighting function based on a correlation-based comparison between the internal representations of both signals. This allows the correlation-based models to be used for other degraded versions of the speech, *e.g.,* speech processed with speech enhancement algorithms, since the distortion, *i.e.,* noise, is not required to be a separate signal in isolation from the clean speech.

A number of extended versions of the SNR-based models were proposed based on

correlation and coherence compression to address their limitations. For instance, Holube and Kollmeier (1996) proposed a temporal-based measure based on the STI known as normalised covariance based STI (CSTI) (Goldsworthy and Greenberg, 2004), sometimes known as normalised covariance metric (NCM). Its main distinction from the STI is the fact that the CSTI used the covariance between octave-band temporal envelopes of the input signals, whereas the STI model used the differences in their modulation transfer functions. The CSTI was shown to correlate well with the intelligibility of vocoded speech Chen and Loizou (2011) that was used to simulate speech processed through cochlear implant. Furthermore, Kates and Arehart (2005) introduced a Coherence Speech Intelligibility Index (CSII) measure as an extension of the original SII to account for nonlinear processing artefacts like peak and centre clipping often occurring in hearing aids (Kates and Arehart, 2005). The essential differences from the SII is that the CSII uses the magnitude squared coherence (MSC) function (Carter et al., 1973), defined as a measure of correlation between complex Fourier-coefficients over time as a function of frequency, to estimate the speech and noise spectra. It is evident that both CSTI and CSII measures examine the temporal correlation within each frequency band.

Additionally, more advanced models were further proposed by Christiansen et al. (2010), Boldt and Ellis (2009), and Taal et al. (2011) by considering the correlation in the joint spectro-temporal domain. In particular, Christiansen et al. (2010) used the advanced auditory model proposed by Dau et al. (1996) as an intelligibility predictor, hence it known as DAU model. The model determines the spectro-temporal internal representations of the input signals as described in Dau et al. (1996), and then segmented in short-time frames within each frequency band. Finally, each frame was compared by means of a correlation coefficient.

Boldt and Ellis (2009) defined a normalised subband envelope correlation (NSEC) using an auditory based filterbank, *i.e.,* gammatone filterbank, to get the internal representations of the input signals. The NSEC was then defined by the normalised correlation over all time and frequency points. The results of DAU and NSEC models were highly correlated with noisy processed speech.

Moreover, Taal et al. (2011) introduced a Short-Time Objective Intelligibility (STOI) measure, denoting a correlation of the temporal envelopes between the reference clean and degraded speech in short-time overlapped segments. The inputs signals were first decomposed by a 1/3-octave filter bank, segmented into short-time windows, normalised, clipped, and eventually compared by means of a correlation coefficient. It was demonstrated to be highly correlated to human speech intelligibility score of spectro-temporal

weighted noisy speech and enhanced speech.

2.4 Microscopic Intelligibility Models

As stated previously, the macroscopic models have been developed to predict the average speech intelligibility based on long-term statistics given the specification of a communication channel. An alternative, termed the 'microscopic' models (Barker and Cooke, 2007), aim to make a more fine-grained level of predictions about listeners' responses to speech. In this section, we describe two approaches to microscopic intelligibility modelling. First, we describe models that are expected to make a predication based on characterising phonetic confusions using a one or more simple fixed template matching. The second approach, named 'statistical classification based models', represents the models that are expected to make a predication based on characterising phonetic confusions using statistical modelling. These approaches are described in Section 2.4.1 and Section 2.4.2, respectively.

2.4.1 Template Matching Based Modelling

This non-parametric approach of microscopic modelling relies on matching the stimulus to a set of stored, pre-defined 'templates' to predict speech intelligibility. Templates here refer to phonetic classes. This approach is often adopted to investigate phoneme misidentifications within acoustically-similar - or between acoustically-different - phonetic classes.

The study by Miller and Nicely (1955) was the first to provide some early insights on perceptual confusions and was particularly studied among consonants. Closed-set phonetic identification tasks were conducted using consonants-vowels (CVs) in white noise and different bandpass filtering conditions and analysed using confusion matrices (CM). In particular, the CVs comprised of 16 most common English consonants followed by the vowel /a/, (*e.g.,* /ba/, /ta/). They observed a number of perceptual confusion groups of consonants (*e.g.,* /p, t, k/), and they further analysed the data by considering the information transmitted through different articulatory features (voicing, nasality, affrication, duration, and place of articulation). Wang and Bilger (1973) conducted an analysis using consonant-vowel combinations (CVs or VCs), which included 25 consonants and the vowels /a, i, u/, and performed an articulatory feature-based analysis. Their findings showed that consonant perception depends on the consonant itself, and

the vowel context that the consonant is embedded in.

Furthermore, Allen (2005b) re-investigated the Miller and Nicely (1955) data and associated them with the Articulation Index (AI). The author suggested that the CM should be evaluated in terms of perceptual events instead of articulatory features. He therefore proposed a model known as the confusion pattern (CP). The CP defined as the proportions of the different response alternatives as a function of the experimental conditions (*e.g.,* SNRs) for a given speech stimulus. It was demonstrated that the CP was more accurate than the CM for identifying perceptual confusion groups because it gave the relative intelligibility of the consonants in CV contexts across experimental conditions.

Régnier and Allen (2008) developed a computational model of speech reception known as the AI-gram by combining the AI model of speech intelligibility and a simple linear auditory model filterbank. The AI-gram representation presented an initial estimate of audibility of various spectro-temporal components for a given speech signal in noise to study how a natural speech signal is decoded by the auditory system. To assess the importance of audible components to speech recognition, as predicted by the AI-gram, a systematic psychoacoustic method known as Three-Dimensional Deep Search (3DDS) approach was developed by Li et al. (2010, 2012). The concept behind the 3DDS is to systematically remove various parts of a speech sound and then to evaluate the significance of the removed component from the change in the recognition score. This was achieved through identifying the spectro-temporal cue regions of consonants using consonant recognition data obtained with noise masking and spectral filtering.

Singh and Allen (2012) used the recognition of stop consonants for modelling the perceptual confusion. They found that there were large differences in the recognition of stop consonants both across consonants and for individual tokens of a given consonant. This was consistent with the findings of Phatak et al. (2008) using consonant recognition and confusions. Moreover, Toscano and Allen (2014) analysed across- and within-consonant recognition errors for CVs composed of the 16 consonants used by Miller and Nicely (1955) study followed by four different vowels. The findings suggested that there was a large variability in the consonant recognition across consonants as well as within consonants. A recent study by Zaar and Dau (2015) investigated the relative importance of the factors that influence consonant perception both in terms of stimulus-related (*i.e.,* source) and listener-related (*i.e.,* receiver) effects using 15 Danish consonants combined with the vowel /i/ as CVs. The results confirmed the large speech-token induced variability of consonant-in-noise perception.

2.4.2 Statistical Classification Based Modelling

The second approach to the microscopic modelling is statistical classification based. Its primary objective is to predict listeners' responses to a specific noisy speech token at a fine-grained level of detail using statistical classification techniques. Most studies of the statistical based models require an auditory processing model that extracts features of a speech signal to be recognised, and an automatic speech recogniser (ASR).

Examples of earlier studies on microscopic modelling that used ASR system to predict responses to individual tokens were done by Ainsworth and Meyer (1994); Ghitza (1993) and Holube and Kollmeier (1996). For instance, Holube and Kollmeier (1996) conducted the first study that focused on microscopic modelling of speech recognition using a rhyme test, *i.e.,* recognition of single meaningful words. Jürgens et al. (2007) and Jürgens and Brand (2009) used a similar approach to Holube and Kollmeier (1996) in which an auditory preprocessing based on the model of Dau et al. (1996) and a dynamic-time-warp (DTW) described by Sakoe and Chiba (1978) speech recogniser were applied. The studies were conducted using German logatomes in a vowel-consonant-vowel and a consonant-vowel-consonant combinations (CVCs and VCVs) (Wesker et al., 2005). Both studies demonstrated that their microscopic model was able to distinguish noisy speech signals in a closed-set testing procedure.

Unlike the previous studies in which a non-parametric DTW speech recogniser were employed, studies by Cooke (2006) and Barker and Cooke (2007) used a parametric statistical modelling techniques. In particular, Cooke (2006) and Barker and Cooke (2007) applied missing data speech recognition techniques (Cooke et al., 2001) to predict the response of listeners to the noisy speech on a token by token basis. The concept behind the missing data speech recognition is to estimate prior to decoding which spectro-temporal elements of the spectro-temporal representations were dominated by speech energy (*i.e.,* reliable) and which were dominated by background noise (*i.e.,* unreliable). The reliability estimates formed a 'missing data mask', this information was used to treat reliable features differently based on the unreliable ones during decoding (Cooke et al., 2001). In particular, Cooke (2006) proposed a 'glimpsing model' in which the glimpses of a signal represented by the connected regions in a spectro-temporal representation greater than a certain minimum area computed from the number of spectro-temporal pixels and where each spectro-temporal pixel had a local SNR larger than a threshold. In particular, the model was assessed in terms of its ability to predict identification rates and confusion matrices for a set of consonants presented in noise. Consonants were presented in a VCV context with a fixed vowel, /a/. The model showed a high correlation

with listeners' performance for both stationary and fluctuating noise. Cooke's glimpsing model was further extended in Barker and Cooke (2007). Barker and Cooke (2007) evaluated the model using a recognition task of spoken letter-digit grid references. Their findings indicated that the model's prediction of intelligibility was highly-correlated with actual intelligibility when the amount of glimpses was low, and less-correlated when the speech signal was less masked. In general, the model was shown most accurate predictions of the intelligibility of individual spoken letters and different talkers in adverse conditions.

We will see in Chapters 6 how the missing data ASR can be used to define a discriminative intelligibility model and thus a detailed description of how to handle missing data will be given then.

2.5 Chapter Summary

So far we have presented the basic ideas behind perceiving speech in noise, and reviewed the literature related to speech intelligibility modelling. We mentioned that the models of intelligibility are broadly classified into macroscopic and microscopic. The main differences between both classes is that the macroscopic models make prediction based on long-term statistic of the inputs signals by either quantifying the SNR or the correlation between the input and a reference signal. Microscopic model, on the other hand, make use of a finer level of precision by characterising characterising phonetic confusions using either a template-matching model or a statistical learning.

As stated before, the main motivation for developing speech intelligibility models was twofold. Firstly, the intelligibility models are of great practical importance to replace costly listening tests in early stages of the development phase. Secondly, the intelligibility models help in guiding the development process of speech enhancement algorithms. We will see in Chapter 4 how the intelligibility model is embedded into a closed-loop near-end enhancement framework. The following chapter is devoted to review algorithms for enhancing speech intelligibility.

Chapter 3

Algorithms for Enhancing Speech Intelligibility

3.1 Introduction

In contrast to conventional noise reduction algorithms, the algorithms reviewed in this chapter aim to enhance the intelligibility of far-end speech for the near-end listener who is located in an environment with background noise, reverberation, or both. Since the acoustical background distortions reach the near-end listener's ears directly and hence can hardly be controlled, a possible solution is to *pre-enhance* the far-end speech before playback in order to become more intelligible in presence of the background. Example application scenarios would be a train-station, in which the intelligibility of an announcement is degraded by a passing train and reverberation, and mobile telephony, where people often make phone calls in challenging acoustical environments in which conversation can be perceptually difficult.

A straightforward solution would be to increase the level of the speech. This approach, however, arrives to a certain point and then the increased in the volume may not be possible anymore due to loudspeaker limitations. Additionally, playback levels may be reached which are uncomfortable or may result in hearing damage over long exposure. A more attractive approach would be to leave the speech energy unchanged but instead *redistribute* energy within the speech signal over time and/or frequency.

We call this approach energy-constrained near-end intelligibility enhancement. A number of speech modification algorithms have been presented in the literature to tackle the problem of near-end intelligibility enhancement. Several names have also emerged for this new direction in speech enhancement which include 'speech intelligibility en-

hancement', 'speech reinforcement', 'near-end listening enhancement', or simply 'speech pre-enhancement'.

The remainder of this chapter is outlined as follows: since talkers naturally modify their speaking styles according to the listening context, *e.g.,* as a response to the near-end additive noise known as 'Lombard effect' (Lombard, 1911), we review the acoustic changes that occur during intelligibility-enhancing speaking styles in Section 3.2. Thereafter, in Section 3.3, we describe the near-end intelligibility enhancement algorithms designed to operate in additive noise. Section 3.4 outline the algorithms that have been developed to enhance speech intelligibility in the presence of reverberation or both additive noise and reverberation, respectively. Finally, the chapter concludes with a general discussion in Section 3.5, followed by a summary in Section 3.6.

3.2 Speech Production in Adverse Environments

Speakers appear to naturally and spontaneously adopt a distinct intelligibility-enhancing style of speech production when they are aware of the difficulty of their immediate communication context. The context might be environment related, *i.e.,* a noisy environment, leading to so-called 'Lombard speech' or it might be related to listeners' perception, *i.e.,* listeners with hearing impairment or non-native listeners, resulting in a so-called 'clear speech'. The adaptations of the intelligibility-enhancing speaking styles may happen at two different levels: (i) at an acoustic level including changes in phonation, place and manner of articulation (Cooke and Lu, 2010; Picheny et al., 1985; Van Summers et al., 1988); or (ii) at a linguistic level, including changes in words and vocabulary (Howell et al., 2006; Lindblom, 1990; Patel and Schell, 2008; Uther et al., 2007). In this section, we summarise findings from studies reporting acoustic-phonetic changes occurring in Lombard and clear speaking styles which result in increased intelligibility. The purpose of this review is to inform the development of the near-end speech intelligibility enhancement strategies which improve robust communication.

3.2.1 Lombard speech

The increase in the vocal effort associated with producing speech in noisy environment is normally referred to as Lombard effect (Lombard, 1911) and the speech style generated in such environment is called Lombard speech. Children and grown-up people usually experienced the need to alter their speaking style in noisy acoustic environment, *e.g.,* a

crowded restaurant, attempting to enhance information transfer to their listener in an involuntary and self-monitoring manner.

Classic Lombard speech studies have proposed several procedures to record Lombard speech that differ in the use of noise type, the noise level, and the nature of the speech task, *i.e.,* a non-communicative or communicative task. Dreher and O'Neill (1957) recruited 15 talkers (3 males and 12 females) to record stimulus materials (read words and sentences) with quiet, noise levels of 70, 80, 90, and 100 dB sound pressure level (SPL). During recording, each talker was asked to wear a pair of headphone which fid with a white noise generator, and then he/she read the materials. Van Summers et al. (1988) employed 2 subjects to read words in quite (at 33 to 37 dB SPL) and at presentation level of 80, 90, and 100 dB SPL. The additive noise was broadband white noise. Pittman and Wiley (2001) used read-list single words (50 target words) presented at quiet and 80 dB SPL in wide band noise and multi-talker babble at 80 dB. Similarly, Lu and Cooke (2008) produced stimulus materials in quiet and in the presence of noise at a number of levels. They recruited 8 talkers whose asked to read out 400 sentences in each of quiet and 3 speech-shaped noise conditions at 82, 89 and 96 dB SPL. These studies belonged the the non-communicative task where there is no interaction with a listener. On the other hand, in the procedure where the communicative task is used, talker-listener pairs seated face to face and communicating word lists in conditions of quiet and noise suggested by Webster and Klumpp (1962). For instance, Junqua et al. (1998, 1999) made a comparison between speech generated when reading a list of phrases with that generated while talking to a voice dialling system at 85 dB SPL. In these studies, 10 talkers (5 males and 5 females) were recorder in quiet and with 3 different types of noise, including pink noise, white noise spectrally shaped, and speech-like noise (according to long term speech spectrum), were used for the Lombard speech recordings. In the study by Cooke and Lu (2010), speech produced by talkers speaking alone or in pairs. Talkers generated speech in quiet and in backgrounds of speech-shaped noise, speech-modulated noise, and competing speech.

Lombard speech has been shown to involve several acoustic changes which includes: an increase in vocal intensity, decrease in speaking rate, higher f_0, greater f_0 range, increase in vowel duration, reduction in spectral tilt (*i.e.,* relative increased energy in the high-frequency components), and increase in the first formant (F1) and second formant (F2) frequencies (Garnier et al., 2006; Hansen, 1996; Junqua, 1993; Lu and Cooke, 2008; Van Summers et al., 1988). A further observation is that the energy shifts from consonant to vowels (Garnier et al., 2006; Junqua, 1993; Womack and Hansen, 1996) and

from semivowels to vowels and consonants (Hansen, 1996). In addition, a recent work by Drugman and Dutoit (2010) demonstrated that the increase in the vocal effort resulted in a significant manipulation in the glottal source through an increase in f_0, reduction in the number of harmonics in the amplitude spectrum, and reduction in the H1-H2 ratio, *i.e.,* the ratio between the amplitude of the glottal spectrum at f_0 and at the second harmonic. The authors also reported boosting the energy of particular spectrum frequency bands that comprise both the glottal and vocal tract changes through increases in E21, *i.e.,*the energy ratio between the frequency band 1-3 kHz and 0-1 kHz, and E31, *i.e.,*the energy ratio between band 3-8 kHz and 0-1 kHz. However, the degree of increase in E21 is significantly higher compared to the increase observed in the E31. A related line of research investigated the loudness of Lombard speech (Godoy and Stylianou, 2012). In particular, Godoy and Stylianou (2012) found that the Lombard speech increased the loudness for voiced speech parts by boosting the average spectral energy in the 500-4500Hz frequency band thus increasing the audibility of the formants. That is, the voiced parts were louder while the unvoiced parts were quiet, this observation was in line with the findings obtained by Junqua (1993) where the energy shifted from consonant to vowels.

The Lombard intelligibility benefits has also been linked to other possible factors which includes: the noise level (Lu and Cooke, 2008; Van Summers et al., 1988), the noise nature, *i.e.,*whether it is an energetic or informational masker (Cooke and Lu, 2010; Lu and Cooke, 2008), the noise spectral content (Lu and Cooke, 2009b), the type of speaking task, *i.e.,*whether it is a read clear speech or conversational speech (Aubanel et al., 2011), the linguistic content (Patel and Schell, 2008), and the presence of visual cues (Fitzpatrick et al., 2011). In particular, Cooke and Lu (2010) showed that the talkers usually benefit from the gaps in the masker when speaking in the presence of amplitude-modulated noise by predicting pauses in the background noise and timing their speech. Patel and Schell (2008) noted that the increase in Lombard intelligibility varied form speaker to speaker and more sensitive to the content of the speaker's speech, *i.e.,* when the noise level is high, the f_0 and duration of words bearing a high informational load are more likely to be modified by the speaker.

What is not so obvious is the extent to which the acoustic changes that talkers adopt relate to specific properties of the noise. In Lu and Cooke (2009b), Lombard speech was collected in the presence of low-pass and high-pass filtered noise. The authors reported a flat spectral tilt for both low-pass filtered noise and high-pass filtered noise. However, in the case of high-pass filtered noise, the flattening was not significant since the energy

is redistributed to the most masked region.

Furthermore, the speech produced in noise has been found to be more intelligible than speech produced in quiet when both speech types presented in noise at the same level of SNR (Dreher and O'Neill, 1957; Junqua, 1993; Lu and Cooke, 2008; Van Summers et al., 1988). Junqua (1996) reported that the type of masking noise and the gender of the talkers used for the experiment played an important role in improving the intelligibility of both types of speech. Junqua (1996) also found that acoustic changes of speech produced in noise are highly talker-dependent since they differ from person to person. Lu and Cooke (2008) found a high intelligibility relative gain of about of 59%. The gain was measured between speech produced in noise-free and in noisy conditions in the presence of speech-shaped masker at -9 dB SNR. However, the acoustic properties responsible for this intelligibility benefit and how they linked to the properties of the noise and the task involved are not entirely clear. One observed factor in the Lombard speech is the changes in f_0 but its relationship to the intelligibility benefit is far less understood. A study by Lu and Cooke (2009a) manipulated natural speech using two factors as observed in Lombard speech which are f_0 and spectral tilt. They found that the modification in spectral tilt played an important role in gaining the intelligibility benefit, whereas the change in f_0 was less effective since no gain obtained when modified in isolation and with further improvement in combination with the spectral tilt modification.

3.2.2 Clear speech

Clear speech refers to a speaking style that a speaker deliberately adopts in order to maximise intelligibility when facing a communication barrier (Picheny et al., 1985; Smiljanić and Bradlow, 2009). Speakers may utilise this style of speaking in several scenarios, for instance, when talking in a noisy environment and when talking to a hearing-impaired, a non-native listener, or to a speech recogniser. In general, clear speech is an extremely articulated speech style that vary from speaker to speaker, and it requires an increased effort so that the listener produce less effort to discriminate between sounds.

Numerous studies have been directed toward examining the characteristic exhibited in clear speech for promoting intelligibility, *e.g.,* (Amano-Kusumoto and Hosom, 2010; Drullman et al., 1994a,b; Hazan and Baker, 2011; Krause and Braida, 2004; Picheny et al., 1985, 1986, 1989; Smiljanić and Bradlow, 2009). These studies have found that clear speech is characterised by: a decrease in speaking rate (associated with an increase in vowel duration and a longer, more frequent pauses), an increase in consonant energy, an increase in f_0, spectral flattening (*i.e.,* an increase in energy at higher frequencies),

an increase in modulation depth in the temporal signal envelope and an expansion in vowel space (with corresponding F1 and F2 shifts). As Amano-Kusumoto and Hosom (2010) stated, spectral flattening and vowel space expansion are among the most effective modifications (vowel space is a widely used acoustic metric to describe articulatory function, and usually represented as a two dimensional F1-F2 space that is bounded by the first two formants of the corner vowels, *i.e.,* /i/, /æ/, /a/, and /u/ (Bradlow et al., 2003; Ferguson and Kewley-Port, 2007)).

A number of authors have considered the role that clear speech plays in enhancing intelligibility for various listener groups. An influential work by Picheny et al. (1985) confirmed the substantial advantage of clear speech intelligibility for hearing-impaired listeners presented with nonsense sentences. These findings were further expanded by Payton et al. (1994) in several degraded conditions including additive noise, reverberation and a combination of noise and reverberation. They reported a significant intelligibility improvement of clear speech for both hearing-impaired and normal-hearing listeners under the tested conditions. In recent years, a large and growing body of literature has confirmed these findings and extended them to other listener populations: adults with normal or impaired hearing (Ferguson and Kewley-Port, 2002; Krause and Braida, 2002; Liu et al., 2004), children with and without learning impairments (Bradlow et al., 2003), native and non-native listeners (Bradlow and Bent, 2002; Smiljanić and Bradlow, 2007) and elderly adults (Helfer, 1998). The effectiveness of clear speech has been demonstrated for audio-only and audio-visual modalities for both younger listeners (Gagne et al., 1994; Gagné et al., 2002) and older listeners (Helfer, 1998). Although clear speech was shown to be beneficial in various communicative situations for many listener populations, as with Lombard speech, the extent to which each observed modification contributes to intelligibility is not yet fully understood (Smiljanić and Bradlow, 2009; Uchanski, 2005).

At first glance, it seems that the Lombard and clear speaking styles are different since the former is seen as a re-action to noise while the latter is seen as a result of explicit instruction to speak clearly. However, there are similarities between the modification changes observed in the two modes of speaking styles including acoustic-phonetic and phonological changes. In fact, the acoustic-phonetic and phonological changes are more likely to exhibit in Lombard speech, as observed in (Junqua, 1993; Lu and Cooke, 2008), thanks to the spectral, durational and other modifications which are not constant across the speech signal and thus vary on a segment-by-segments basis of speech signal. In addition, Bond and Moore (1994) and Cooke and Lu (2010) reported a vowel

space modifications in Lombard speech. The property of vowel space expansion is commonly seen in clear speech and has been demonstrated to promote the intelligibility and therefore benefiting the listeners (Bradlow et al., 1996).

3.3 Near-end Intelligibility Enhancement in Additive Noise

In the previous section, we described the acoustic changes of Lombard and clear speech and their effect on speech intelligibility. We now consider emulating these observed acoustic modifications by applying signal processing techniques to speech produced in quiet.

There is an extensive literature on algorithms aiming at enhancing the intelligibility of speech in noise. There are two types of systems. First, those that modify natural speech signal. Second, those that modify synthetic speech signal by adapting the speech production stages of text-to-speech (TTS) systems, examples included the work by Langner and Black (2005); Raitio et al. (2011) and Valentini-Botinhao et al. (2012). The former is the focus of this thesis, and thus the modification designed for synthetic speech is excluded from this review.

We classified these algorithms into noise-independent and noise-adaptive. Noise-independent algorithms do not take into account any prior information about the noise background. These algorithms are presented in Section 3.3.1. Noise-adaptive algorithms utilise prior knowledge or estimates of the background noise and adapt their processing accordingly. Section 3.3.2 presents an overview of these noise-adaptive speech modification algorithms.

3.3.1 Noise-independent Algorithms

Algorithms in this category generally work by mimicking the acoustic changes observed in studies of speech produced in noise. In this section, we survey the literature in regards to the near-end intelligibility enhancement which are noise-independent.

Boosting the Consonant-Vowel Power Ratio

Boosting the consonant-vowel ratio, an effect commonly seen in clear speech, was among the first methods proposed for enhancing the intelligibility of speech produced in noise

(Kretsinger and Young, 1960). It concentrated mainly on giving more weight to those speech units that have been associated with the increase in intelligibility. Therefore, low in amplitude speech units, *i.e.,* normally consonants, are reinforced with respect to the stronger speech units, *i.e.,* normally vowels. In general, this approach includes methods such as amplitude equalisation (sometimes referred to as dynamic amplitude compression) or a system of amplitude equalisation combined with high-pass filtering in order to boost the second formant frequencies in respect to the first formant.

The dynamic amplitude compression typically refers the the process of reducing the amount of gain applied to a signal when the input increases above a specified threshold level in an automatic way. However, if a large amount of gain is applied the signal becomes clipped in either analogue or digital domains. Clipping introduces a large amount of undesirable harmonic distortion due to its non-linear nature (Kretsinger and Young, 1960). As an alternative to peak clipping, the authors proposed a compression limiting technique where the compression is only active at high levels resulting in a high compression ratio. They found that using the compression limiting was substantially better than the peak clipping.

Thomas and Niederjohn (1968, 1970) suggested using a highpass filtering in combination with infinite amplitude clipping. The infinite amplitude clipping derived by mapping all positive and negative values to their equivalent of maximum positive and negative amplitude respectively and then ends up with a binary time-domain signal. They reported that the intelligibility of the bandpass filtered speech in white noise yielded significant improvement by up to 50 % at 0 dB SNR. Nevertheless, the low threshold (just below the clipping level) and the high compression ratio introduced by clipping result in the severe distortion which may cancel the effectiveness of increasing the power of consonants. To alleviate this shortcoming, Niederjohn and Grotelueschen (1976, 1978) developed a rapid amplitude compression. The results of intelligibility gain obtained by applying this techniques was identical to the findings of Thomas and Niederjohn (1970) at 0 dB SNR, and was significantly higher at lower SNRs.

Skowronski and Harris (2006) suggested an automatic voicing detector to redistribute the energy from voiced regions to unvoiced regions. The unvoiced regions of the speech signal were detected using a simple spectral flatness measure and boosted by a ratio of 7.4 dB. The ratio was chosen based on empirical considerations. The investigators showed intelligibility increases of up to 15% for words presented by a range of talkers in the presence of white noise at 0 and -10 dB SNRs. However, there was found to be a large variability across sounds and speakers.

Empirically, Furui (1986) observed that transient components in speech are more essential than stationary ones. This observation has led to algorithms for emphasising transients. For instance, Yoo et al. (2007) utilised a high-pass filter followed by a signal decomposition into quasi-steady-state and transient components. They amplified the latter components, including transitions between vowels and consonants and within vowels, with respect to the former by an empirically determined factor. They found that the modified speech yielded a higher intelligibility gain in comparison to the original speech presented in speech-weighted noise at -10 dB SNR. Tantibundhit et al. (2007) extended the notion of the transient's amplification by decomposing speech into tonal, transient, and residual components using a hidden Markov chain based on a modified discrete cosine transform and a wavelet-based hidden Markov tree. Similar to the work by Yoo et al. (2007), the transient components are amplified by a heuristically determined amount and mixed to the original speech, and finally the energy is re-normalised. Rasetshwane et al. (2009) implemented a wavelet packet-based technique in order to extract an estimate of the transient speech components. The estimated transient components were then amplified and recombined with the original speech. The results of this technique using modified rhyme tests indicated its importance in improving the intelligibility in line with the findings obtained in both (Yoo et al., 2007) and (Tantibundhit et al., 2007).

Chanda and Park (2007) introduced a low-complexity system that applied a tunable bandpass filter to emphasise consonants relative to vowels. To ensure that the input level remains roughly the same as the output level, the authors dynamically adjusted the cut-off frequency of the filter. Their objective evaluation demonstrated that modified speech, obtained using this technique, gave significantly improved SII scores, and particularly for male talkers.

Spectral Tilt Flattening and Formant Enhancement

The algorithms here are based on the empirical considerations that high frequencies are crucial for improving speech intelligibility. In Thomas (1968) and Thomas and Ohley (1972), a highpass filtering was implemented to emphasise the F2 formant relative to F1. McLoughlin and Chance (1997) used line spectral pairs in which each formant is shifted upwards in frequency to improve a so-called 'formant-to-noise' ratio. They further flatten the spectral tilt by widening the formant bandwidth. Preliminary results of an informal listening test showed that using the formant shift in isolation increased speech intelligibility by 10 % and using formant bandwidth adjustment in isolation improved

the intelligibility by 14 %.

Hall and Flanagan (2010) used differentiation, *i.e.,* a first-order backward difference, and formant equalisation. Both techniques involving a high-pass type filter. They compared the effects of these techniques on the intelligibility of telephone speech and inferred that both differentiation and formant equalisation produced intelligibility improvement over original speech.

Jokinen et al. (2012) made a comparison between two straightforward post-filters, that share the principle of transferring energy from the F1 to higher frequencies, with the formant equalisation proposed by Hall and Flanagan (2010). The aim of former post-filter was to adaptively track the formant locations, whereas the objective of the latter one was to use fixed locations. A subjective evaluation using a Speech Reception Threshold (SRT) showed both the post-filtering techniques resulted in an intelligibility improvement compared to original speech. However, there was not a substantial advantage of one technique over the other.

Very recently, Koutsogiannaki and Stylianou (2014) proposed a so-called 'mix-filtering' to imitate the acoustic properties of clear speech and thus enhance the causal speech in the presence of noise. The mix-filtering utilises a multi-band filtering scheme to extract the information of the importance frequency bands and then, combines this information with the original signal. Their findings indicated the effectiveness of the mix-filtering technique in improving the intelligibility of casual speech while maintaining its quality.

Modification of Duration and Prosody

Huang et al. (2010) attempted to mimic the Lombard effect by modifying phoneme duration, f_0, formant frequencies, formant bandwidth and energy in each frequency band (*i.e.,* spectral envelope). The modification was carried out using STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (Kawahara et al., 1999)). Huang et al. (2010) assessed this approach by comparing the synthesised Lombard speech with noise-free Lombard speech in regards to their similarity, naturalness and voice quality. They found that the proposed system was able to manipulate the original speech into a synthetic Lombard speech which was equivalent to the natural Lombard speech in term of quality.

Pitch and Temporal Envelope Enhancement

Park et al. (2010) proposed to enhance the fluctuation of the band-wise temporal envelope and reinforce the pitch structure in frequency domain. Code excited linear prediction-based (CELP) with a pitch-enhancing postfilter was implemented in order to control pitch enhancement in fine spectral resolution. The authors used a single pitch period in the pitch-enhancing postfilter. The pitch period is determined in time domain using an autocorrelation technique and then converted it to the pitch frequency. In their proposed algorithm, the spectrum is divided into non-overlapping pitch bands. Then the relative range between the harmonic peak and valley of each pitch band is expanded while maintaining the harmonic shape. Park et al. (2010) found that harmonic valleys below 2 kHz yield much reduction compared to harmonic peaks in which it maintain the original shapes of modified male speech using the pitch enhancement algorithm. A formal listening test demonstrated that the technique enhanced the perceptual quality of speech in several noisy conditions.

Formant and Loudness Enhancement

More recently, Zorila et al. (2012) introduced a system with spectral shaping and dynamic range compression. The system preforms a spectral enhancement (spectral shaping) followed by temporal enhancement (dynamic range compression) which are combined in a cascaded way to form the proposed system. The spectral shaping system consists of two cascaded subsystems which are adaptive and fixed spectral shaping. The adaptive spectral Shaping is adapted to the probability of speech frame voicing and it applies formant enhancement. In the fixed spectral shaping, adaptive pre-emphasis filter is used in oder to control the distortion of speech modification by limiting the attenuation of high frequencies in speech signal. The output of the spectral shaping system is then input to the dynamic range compression. During the dynamic compression, the envelope of the total time signal is dynamically compressed with a 2 ms time constant using a moving average threshold with order determined by the average pitch of speaker's gender. The purpose of this this sub-system is to move the energy from higher region to lower ones. The results of applying the system yielded good intelligibility improvements compared to original speech.

3.3.2 Noise-adaptive Algorithms

In addition to the algorithms described in the preceding section, there is another class of algorithms that exploit the characteristics of the near-end noise or the noise statistics. This class is referred to noise-adaptive. In this section, we give an overview of the available noise-adaptive algorithms in the literature.

Formant Enhancement

Formant enhancement was implemented by Brouckxon et al. (2008) based on the masking effects of the auditory system in order to enhance the first three formants. Precisely, the authors suggested to compute the signal-to-masking ratio between the current speech sound pressure level (SPL) and the hearing threshold based on the instantaneous background noise on a formant-by-formant basis. Their results showed that this approach produced an increase in the intelligibility by around 4 dB lower SRT.

Modification of Local SNR Ratio

The fundamental idea of this approach is to reallocate energy by modifying the local SNR. Sauert and Vary (2006b) proposed a time adaptive and frequency dependent amplification of the speech signal to reestablish the distance between the average measured speech spectrum and the average measured noise spectrum and thus recovering a certain local target SNR. Although this algorithm improved the intelligibility of the speech in the presence of background noise, it increased the energy of speech signal. Similarly, (Sauert and Vary, 2006a) proposed SNR recovery algorithm based on a frequency dependent filtering of the speech signal by amplifying some frequency components and attenuates others using two possible strategies. The first strategy is referred to as 'equal SNR' and its principle is to reduce speech energy in frequency bands where the speech energy is higher than the noise, *i.e.,*with high SNR, and to amplify speech energy in frequency bands where the speech is covered by the noise, *i.e.,* with low SNR. It ended up with an equal SNR at all frequencies. The second strategy named as 'maximal power transfer' was motivated by a simple model of human hearing. It finds the most masked frequency bands and decrease the speech energy and distributed it back among clean frequency bands. Objective evaluation showed that the latter algorithm increased intelligibility better compared to the original speech, whereas the algorithm of equal SNR did not improve the intelligibility estimate over almost the whole range of tested SNR except at high SNRs where a slight increase can be found. Sauert et al. (2008) further

extended the SNR recovery algorithm with a non-uniform low delay filterbank. The spectral weights were derived for approximately Bark-scaled frequency bands to account for the non-uniform frequency resolution of the human ear.

Tang and Cooke (2010, 2011) investigated several strategies for energy reallocation implemented over time frames or frequency bands or both. In Tang and Cooke (2010), the approaches were grouped into two categories. The first three approaches were based on equalising local SNRs to a fixed global SNR for each time frame, for each frequency band, and for each spectro-temporal element, respectively. In the first approach, a time-varying gain was applied to the speech signal which resulted in constant frame-wise SNR across the signal. The second implemented a time-invariant spectral tilt to equalise the SNR in each frequency band. Further, the third approach was a generalisation of the first and second approaches and modified the local SNR for each spectro-temporal element of speech. The latter approach was applied to a selected frequency bands or a refinement to those selected frequency bands. The first in this category, operated through increasing speech level in selected frequency bands used a fixed pre-determined increase equivalent to 20 dB for the range of bands corresponds to centre frequencies from 1.8-7.5 kHz. The second approach extended the previous one by boosting selected frequency regions whose local SNR is very low. The approaches were evaluated objectively in Tang and Cooke (2010) and subjectively in Tang and Cooke (2011). The results demonstrated that boosting selected frequency bands significantly enhances the intelligibility. It had also been shown to boost the speech energy in those spectro-temporal elements above 1.8 kHz with a local SNR of less than 5 dB. Additionally, Tang and Cooke (2011) studied the implication of pauses insertion at word boundaries on improving speech intelligibility to minimise intense masker epochs, under a constant-duration constraint which led to an increase in speech rate in order to accommodate pause insertion. This technique has resulted in a significant reduction in speech intelligibility because of a reduction in predictability of word boundaries in noise.

Recovery of Partial Loudness

Sauert and Vary (2006b) proposed the SNR recovery algorithms that place the power spectrum of enhanced speech at a certain fixed log distance from the power spectrum of noise under an energy preservation constraint. Similarly, Shin et al. (2009, 2007) proposed using the perceptual loudness in their recovery algorithms that applied in each frequency band. This technique was motivated by the idea that, in noisy environments, the target signal is partially masked and hence delivered with a reduced loudness. Such

effect was simulated using the model of loudness (Moore et al., 1997). Moore et al. (1997) model calculates the specific loudness of the original noise-free signal and partial specific loudness of the noise-distorted signal. The specific loudness refers to the loudness per equivalent rectangular bandwidth (ERB). That means one can compute the loudness by integrating the specific loudness over all ERBs. Additionally, the partial specific loudness is the partial loudness per ERB.

In particular, Shin et al. (2009, 2007) suggested preserving the loudness of the speech signal despite the noise. The technique is performed by amplifying the input speech signal in such a way that the partial specific loudness for each frequency band of the amplified speech in noise becomes equal to that of the original noise-free signal. This technique, however, required increasing the power of the speech signal. The results demonstrated that the amplified speech improved the perceived quality of the noisy speech signal.

Optimisation of Objective Intelligibility Measures

Very recently, works have emerged that develop near-end intelligibility enhancement algorithms by optimising objective models of speech intelligibility. The glimpse proportion (GP) measure (Cooke, 2006) was optimised by Tang and Cooke (2012) and Aubanel and Cooke (2013). In addition, the speech intelligibility index (SII) (ANSI, 1997) was optimised in (Sauert and Vary, 2009, 2010a; Taal et al., 2013). Furthermore, Taal et al. (2012b, 2014) used a perceptual distortion measure based on a spectro-temporal auditory model, known as Short-Term Objective Intelligibility (STOI) (Taal et al., 2012a), to optimally redistribute speech energy over frequency and time. In Petkov et al. (2013), the probability of correct recognition of the transmitted message was optimised.

Tang and Cooke (2012) proposed stationary spectral modifications to reallocate speech energy across frequency bands under globally-constant energy and duration preservation constraints. The spectral-dependent weighting were selected based on a genetic algorithm optimisation technique (Holland, 1975), commonly used for complex, discontinuous or high-dimensional spaces, and the glimpse proportion (Cooke, 2006). The optimisation was preformed offline for different noise types at a range of SNRs whereas during application, *i.e.,* online, a relatively high-level estimate of the noise context *e.g.,* estimation of the noise type and overall SNR. A surprising observation was the consistent discovery of sparse, highly-selective spectral energy weightings, precisely as noise level increases. The subjective findings illustrated higher intelligibility gain for both stationary and highly non-stationary maskers at a range of SNRs.

Aubanel and Cooke (2013) proposed a temporal expansion by optimising the GP measure to shift potentially-informative parts of the signal to regions where they would be less distorted by noise in the presence of a fluctuating masker. The algorithm defined an alignment between speech and maker for pairs of time frames and perform the expansion accordingly. The objective and subjective evaluation showed the intelligibility improvement of this algorithm especially at lower SNRs. The effect of pause insertion by modifying local speech rate to minimise overlap with a known fluctuating masker was also investigated and found a significant *reduction* in intelligibility which is possibly due to the disruption in a listener's ability to detect the locations of words presented under masked conditions.

Taal et al. (2012b, 2014) recently introduced a way of redistributing the speech energy over time and frequency by optimising a perceptual distortion measure, which is based on a spectro-temporal auditory model, *i.e.,* STOI (Taal et al., 2012a). Taal et al. (2012b, 2014) claim that the STOI measure takes into account short-time information compared to spectral-only models implemented by Sauert et al. (2006); Sauert and Vary (2010a). Thus the proposed algorithms becomes more sensitive to transient regions, which accordingly receive more amplification with respect to stationary vowels. This is similar to the approach of amplifying the consonant-vowel ratio described in the noise-independent algorithms that has shown its effectiveness in enhancing the intelligibility of speech. In Taal et al. (2012b) an objective intelligibility evaluation was preformed and the results predicted that this method yielded higher intelligibility gain in stationary and non-stationary noise conditions. Additionally, in Taal et al. (2014) a subjective listening test was performed and results indicated a substantial increase in intelligibility while still preserving quality.

Taal et al. (2013) introduced a new linear filter by redistributing the speech energy over frequency bands such that an approximation of the SII was maximised. SII predictions and intelligibility listening test demonstrated a significant intelligibility improvements. Sauert and Vary (2010b) proposed a time- and frequency-dependent amplification algorithm by optimising the SII, in a recursive closed-form optimisation. In the recursive closed-form optimisation scheme, the time-varying gain was computed, for each time frame, that maximised the SII given the current noise spectrum levels under the constraint of unchanged short-term audio power of the speech signal. For this reason, a (warped) filterbank with non-uniform frequency resolution was used to divide speech and noise signal into 21 approximately Bark-scaled band signals. The reason of choosing the SII is the spectral weighting function known as band-importance functions which

indicates the degree of speech intelligibility in the spectral domain by estimating the effective amount of audible information in each frequency band based on the speech and noise spectrum level. The objective evaluation by means of the average SII demonstrated large intelligibility improvement compared to previously proposed in (Sauert and Vary, 2009) which was SNR recovery algorithm by optimising the SII. Although an increase in the SII was observed in (Sauert and Vary, 2010b), the spectral adaptation to the noise characteristics was not achieved by this algorithm. This is unnecessary for noises with band-pass characteristics. It particularly occurred when the speech signal after applying the algorithm, *i.e.,*processed speech, has similar band-pass characteristics as the noise and is therefore highly-distorted. Sauert and Vary (2012) further proposed to use a transition between an SII-based weighting proposed in Sauert and Vary (2010b) and unity-weighting to alleviate the mentioned shortcoming.

Petkov et al. (2013) proposed to use the probability of correct recognition of the transmitted message (at the level of a phonetic transcription) as a measure of speech intelligibility. They optimised the proposed measure for the parameters of two distinct speech modification strategies under an energy-preservation constraint. The two considered modifications were: (i) gain adjustment of band-energies in the bands of an auditory filter-bank; and (ii) gain adjustment of phone-energies. It should be noted that the two modifications are not time scale modification. In addition to the noise statistics, the algorithm required prior knowledge of acoustic speech models (from an automatic speech recognition system) and a transcription of the transmitted message. Furthermore, due to it computational complexity, the algorithm can be implemented in on-line applications. The subjective results showed a higher intelligibility gain compared to both original speech and a reference system by Taal et al. (2012b).

3.3.3 Summary

In this section, we reviewed a large variety of noise-independent and -adaptive near-end intelligibility enhancement algorithms. Recently, Cooke et al. (2013b) compared a large number of these algorithms in a subjective listening test. They reported an improvement in intelligibility of all algorithms that modified the clean speech signal although they have different processing strategies, *i.e.,* spectral shaping and/or dynamic range compression, for a stationary speech-shaped masker under equal energy constraint. For a non-stationary speech masker, the dynamic range compression algorithm developed by Zorila et al. (2012) and Lombard speech reported higher intelligibility improvement, while other algorithms were not able to enhance the intelligibility. The study was then

extended by Cooke et al. (2013a) in the so-called 'Hurricane Challenge'. In the challenge, the noise and SNRs conditions were equivalent to their previous study but more algorithms were evaluated. Findings from the hurricane challenge showed that algorithms that used a dynamic range compression stage resulted in significant intelligibility improvements for the stationary and non-stationary maskers. However, for stationary maskers most gain amplification algorithms were able to improve speech intelligibility (Cooke et al., 2013a).

3.4 Near-end Intelligibility Enhancement in Convolutional Distortions

The algorithms described in the previous section mainly concern the case where near-end intelligibility enhancement is applied to promote intelligibility in the presence of a noisy environment without considering the degradation caused by reverberation. In this section, we review algorithms designed by taking into account the effects of only reverberation on intelligibility and pre-enhancing the speech accordingly without paying attention to additive noise corruptions, or both additive noise and reverberation (which is a very recent research direction), respectively.

3.4.1 Near-end Intelligibility Enhancement in Reverberation

It is essential to note that although the reverberation can reduce the intelligibility of speech, there are some properties of reverberation that can also enhance the intelligibility (Bradley et al., 2003). Room reverberation generates early and late reflections of the signal. The early reflections tend to improve speech intelligibility (Haas, 1972). The late reflections, on the other hand, degrade the intelligibility by filling the gaps in the temporal envelope of speech (overlap-masking) and reducing the low-frequency envelope modulations known for their importance for speech intelligibility (Bolt and MacDonald, 1949; Knudsen, 1929; Nábělek et al., 1989).

Late reverberation decreases the intelligibility of speech due to two types of masking: overlap-masking and self-masking (Nábělek et al., 1989). Overlap-masking occurs when the energy of a phoneme masks the phonemes that follow. As a consequences, the phoneme following the reverberating phoneme becomes difficult to hear. This is more likely to occur when the reverberating phoneme has more energy, *i.e.,* a vowel, and the subsequent phonemes have less energy, *i.e.,* consonants (Arai et al., 2002; Nábělek et al.,

1989). The overlap-masking is the main cause of degradation during reverberation (Bolt and MacDonald, 1949; Nábělek et al., 1989). Self-masking, on the other hand, occurs as a results of the internal temporal smearing of energy within each consonant possibly because of flattened formant transitions (Nábělek et al., 1989).

In a very early study for speech pre-enhancement in a reverberant conditions, Langhans and Strube (1982) attempted to enhance the speech signal by filtering the temporal envelopes of critical bands of the input speech signal using a modulation transfer function. In their work, the temporal envelope filtering was applied as a pre-processing and post-processing based method, however no convincing results are reported. More recently, in a series of papers, (Hodoshima et al., 2002; Kitamura et al., 2000; Kusumoto et al., 2005, 2000), the authors developed the temporal envelope filtering, referred to as modulation filtering, by means of emphasising the important spectral components of the modulation spectrum of speech signal prior to distortion by reverberation (Hodoshima et al., 2002). In particular, the modulation index decrease in reverberant conditions compared to clean conditions where the peak of modulation frequency is around 4 Hz. The important component of the modulation spectrum in clean condition lies between 1 Hz and 16 Hz (especially 2 Hz to 8 Hz) (Hodoshima et al., 2002; Kusumoto et al., 2000).

In the modulation filtering proposed by Kusumoto et al. (2005), two different filter techniques are applied; namely empirically-designed and data-derived filters. The empirically-designed filter enhances components around 8 Hz with a peak at 4 Hz in order to get close to the shape of the original modulation spectrum in reverberant conditions, whereas the data-derived filter recovers the original modulations based on the modulation transfer function estimated from modulated clean and reverberant speech (Avendano and Hermansky, 1996). The data-derived filters are defined as the ratio of the modulation frequency response of clean and corrupted reverberant speech, averaged over a large number of speech signals. The data-derived filter is dependent on the reverberation condition because the filters need to be designed for each environment. Whereas the same modulation filters are applied independent of the environment using the empirically-designed filter.

Arai et al. (2004) and Kusumoto et al. (2005) conducted subjective listening tests with normal hearing and hearing impaired listeners in three reverberant conditions. They observed no significant improvement with normal hearing listeners, however, improved the understanding of reverberant speech for the hearing impaired group.

A further development of the temporal envelope filtering using non-linear processing is called steady-state suppression (Arai et al., 2010, 2002). Steady-state suppression

mainly aims to reduce the amount of overlap-masking. The key concept in the steady-sate suppression, and in modulation filtering, is to emphasise the important temporal dynamics of the speech signal prior to distortion by reverberation. The main difference is that the linear filters are applied to the temporal envelope of speech signals in modulation filtering, while non-linear processing is applied in steady-sate suppression (Arai et al., 2004). In particular, steady-sate suppression aims to reduce the amount of overlap-masking by suppressing the steady-state portions of the speech signal.

In the steady-state suppression technique, the speech signal is first filtered into a number of frequency bands. In each band the envelope is extracted. After down-sampling, the regression coefficients are calculated from the five adjacent values of the time trajectory of the logarithmic envelope as proposed by Furui (1986) to measure the spectral transition and in order to define a speech portion as steady-state. Then the mean square of the regression coefficients are calculated. After up-sampling, the steady-states are calculated when the mean square of the regression coefficients is less than a certain threshold. After determining the steady-state, the amplitude of the portion is multiplied by a factor less than 1.0. The enhanced speech signal then re-synthesised by summing up all the processed signals from each band (Arai et al., 2002). Listening test have shown that the steady-state suppression technique is an effective approach for young listeners and elderly listeners, including both those with normal hearing.

More recently, Koutsogiannaki et al. (2015) proposed spectral and time domain modifications to increase the intelligibility of casual speech in reverberant environments by simulating the acoustic properties of clear speech in terms of spectral energy distribution. A simple spectral transformation, known as mix-filtering, was applied to boost higher spectral regions as appearing on clear speech and taking spectral energy from low-frequency energy which is normally responsible for the overlap masking. This mix-filtering technique is similar to the steady-state suppression techniques to reduce steady-state portions of speech regions and to increase transient information. It also gives a similar acoustic result as steady state suppression. In the time-domain, two techniques for time-scaling casual speech were examined: (i) uniform time-scaling; and (ii) pause insertion and phoneme elongation based on loudness and modulation criteria. Their findings indicated that the combination of spectral transformation and uniform time-scaling gave a significant performance in promoting the intelligibility of casual speech.

3.4.2 Near-end Intelligibility Enhancement in Additive Noise and Reverberation

A large number of algorithms have taken into consideration either a purely additive noise channel or a reverberation-only channel during the development phase. Near-end intelligibility enhancement explicitly targeting reverberant noisy channels has rarely been considered in the literature despite the fact that this is the requirement for many application scenarios. Very recently, Crespo and Hendriks (2014) proposed a time-frequency weighting algorithm for noisy and reverberant environments that works by optimising a slightly modified version of the STOI, and under an energy preservation constraint per frequency band across a segment of time-frames. The masker signal was used to model additive noise and late reverberation, and thus the modified version of STOI minimised the detectability of noise and late reverberation under early speech and locally optimised for each spectro-temporal elements. In principle, the algorithm works in a similar way as the dynamic range compressor, and smears out the energy of the clean speech along time. Objective evaluation of the algorithm indicated its effectiveness in outperforming two reference algorithms, which were the steady state suppressor of Hodoshima et al. (2006), the normalised SNR recovery approach of Sauert et al. (2006), in stationary noise. Crespo and Hendriks (2014) also found that better performance was achieved with slower dynamic range compression, up to the point where overlap-masking occurred during processing.

3.5 Discussion

To summarise the approaches taken by the near-end intelligibility enhancement algorithms from a higher level of abstraction, one could describe the algorithm as an open- or closed-loop system. The open-loop system is usually represented as multiple cascaded subsystems in series or just a single system with an input and output signals. It demonstrates a linear path from the input speech signal to the output speech signal with no feedback loop. A successful example of approach is the spectral shaping and dynamic range comparison developed by Zorila et al. (2012) which does not require a priori knowledge of the noisy acoustic conditions apart from the clean speech signal. Closed-loop systems offer the opportunity to accurately control the process by monitoring the intelligibility estimate of its output signal and *feeding* parameters back to *optimise* the intelligibility estimate under the constraint of having the energy of the output signal of the

system the same as the original desired level of energy power. The latter approach covers the algorithms based on optimisation of objective intelligibility measures, described earlier in this chapter, aiming at measuring, monitoring, and controlling the enhancement process. The measure of the output signal is an objective intelligibility measure (reader refers to Chapter 2 for more information about intelligibility modelling).

We will see in Chapters 5, 6 and 7, how we use the closed-loop feedback system in our new near-end intelligibility enhancement systems. Additionally, we will see the effect of using a combined system of closed- and open-loop system on enhancing the intelligibility of speech in Chapter 7.

3.6 Chapter Summary

In this chapter, we highlighted two intelligibility-enhancing speaking styles that human talkers adapt in several listening scenarios namely: Lombard speech and clear speech. We also surveyed the literature in terms of their acoustic changes.

Furthermore, we reviewed various algorithms for near-end intelligibility enhancement inspired by the acoustic changes observed in the Lombard and clear speech. The algorithms were classified according to the acoustical background for which they have been designed, *i.e.,* for additive noise, reverberation, or both. For the additive noise environments, the algorithms were further categorised based on their noise dependency during the processing which are noise-independent or -adaptive algorithms. We described noise-independent algorithms used in the literature which include: boosting of the consonant-vowel-ratio, flattening the spectral tilt, enhancing the formants or loudness, manipulating duration and prosody, and manipulating the pitch and temporal envelope. Additionally, algorithms that utilised prior knowledge or estimates of the noise context were also reviewed which includes: enhancing the formants, modifying the local SNR, recovering the (partial) loudness of the speech signal, and enhancing intelligibility by optimising an objective criterion.

We stated in Chapter 1 that we aim to exploit information about the talker, thus using the closed-loop feedback system will be more appropriate in this scenario. For this reason, we focus in this thesis on the optimisation based approach for enhancing speech intelligibility. In the following chapter, we will provide an in-depth description of the closed-form optimisation framework, that is used throughout this thesis.

Chapter 4

An Analysis-resynthesis Framework for Pre-enhancement

4.1 Introduction

Generally, to *pre-enhance* a speech signal, one could use an open-loop approach or a closed-loop feedback approach, readers are referred to Chapter 3 for a review of example systems of both approaches.

The main characteristics of the open-loop system can be outlined as: First, each input setting to the enhancement process is fixed (which may or may not exploit information about the acoustic background noise). Second, the system is an open ended non-feedback system which has no control action over the output value.

In the closed-loop feedback system, the parameters of the enhancement process are automatically adjusted based on the current estimate of the objective intelligibility measure. Despite the simplicity in applying the open-loop system, the closed-loop system provides many advantages. First, there is an opportunity to get better parameter estimates for each enhancement iteration without opening the loop. Second, there is an opportunity to exploit external knowledge such as acoustic background and/or information about speakers during the modification process. Finally, there is an opportunity to use a more sophisticated modification design procedure that may require a large amount of computation. For these reasons, we adopt the closed-loop feedback system in this thesis for developing new near-end intelligibility enhancement systems.

In this chapter, we describe a general closed-loop framework for near-end intelligibility enhancement. Figure 4.1 illustrates this framework where the parameters of a modification strategy are automatically adjusted based on maximising an intelligibility

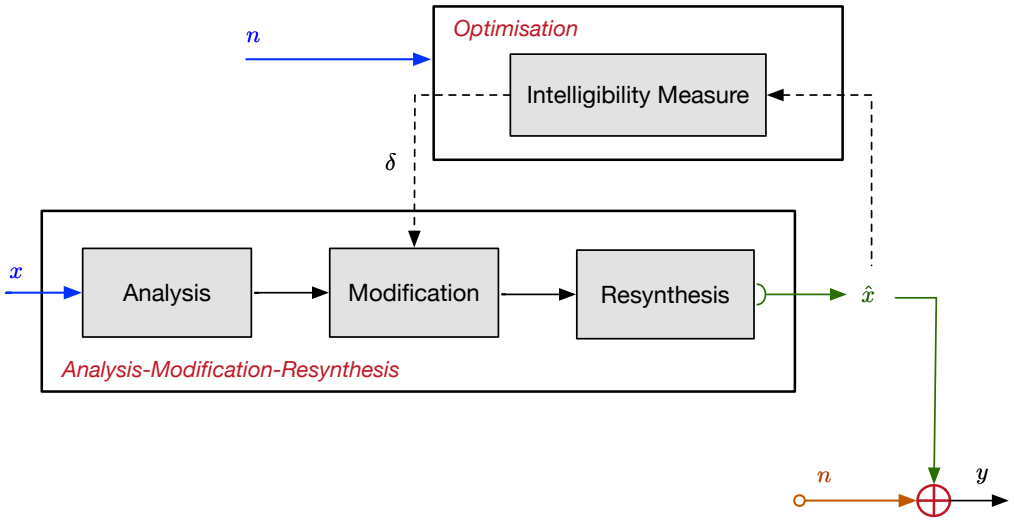

Figure 4.1: A framework for closed-loop intelligibility enhancement system under an energy preservation constraint, where x represents a clean speech signal, n is the environmental noise, δ is the optimised parameters, $\hat{x}$ is the *pre-enhanced* version of speech and y is distorted version after enhancement.

measure, given speech and noise signals, and under an energy preservation constraint.

The organisation of this chapter will follow the presentation illustrated in Figure 4.1. Section 4.2 describes how the incoming speech signal is analysed, and then resynthesised back to time domain to generate a waveform. It should be noted that this chapter is not devoted to give an in-depth history or overview of the analysis-resynthesis, it is rather devoted to give an insight into the analysis-resynthesis used throughout this thesis based on the work by Hohmann (2002). Section 4.3 highlights some possible speech modification strategies. Section 4.4 shows how optimisation can be applied to approach the optimum solution for enhancing the intelligibility. Finally, the chapter ends with section 4.5.

4.2 Analysis-resynthesis

The speech analysis, with its corresponding resynthesis counterpart, are two fundamental components for the closed-loop system, as shown in Figure 4.1. The speech analysis takes the speech signal as its input and uses it to extract a set of parameters that represent the acoustic properties observed from the input speech signal. These parameters are then transmitted to the speech resynthesis component to yield an audible speech signal. In practice, the two components are not disjoint systems that works in a separate fashion,

but rather work together.

For an analysis-resynthesis system to be successful, one should take into account the following: First, the analysis subsystem should allow various independent representations of the features in order to be modified independently and without introducing artifacts-effects. Second, the speech resynthesis should reconstruct a high quality speech signal using the manipulated speech representations.

Many speech analysis-resynthesis methods have been proposed in the literature. Arguably, the most commonly used are channel vocoder (*e.g.,* (Dudley, 1939; Gold and Rader, 1967)), in which a bank of bandpass filters is typically used. Further class of methods is phase vocoder (*e.g.,* (Dolson, 1986; Flanagan and Golden, 1966; Laroche and Dolson, 1999)) which is similar to the channel vocoder. However, instead of considering the amplitude component , the phase vocoder estimates the phase derivative at the output of each filter. Another family of vocoder is named Linear Predictive Coding (LPC) in Atal and Hanauer (1971) that estimates parameters of an all-pole model of the vocal tract.

A more advance channel vocoder techniques is motivated by source-filter theory or auditory processing in humans. They are typically a high-quality vocoder and allow flexible speech modification, since the extracted features are mutually independent. Examples include STRAIGHT which stands for 'Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum' by Kawahara et al. (1999), and auditory filterbank based analysis-resynthesis by Hohmann (2002).

The aim of this section is to explain the theoretical basis for the analysis-resynthesis used in this thesis following the presentation in Hohmann (2002). The two components will be explained individually in 4.2.1 and 4.2.2, respectively.

4.2.1 Analysis

Speech analysis forms the first stage of analysis-resynthesis system architecture. It is a technique that usually used for transforming speech signal $x(t)$ " into a set of parameters, or more generally into a set of signals" Deng and O'Shaughnessy (2003). This section describes an auditory model based speech analysis and how to obtain the features from a speech signal.

In the auditory filterbank based speech analysis, the input speech signal is simultaneously passed through a number of digital filters. Each of which analyses a different range of frequencies, resulting in a narrowband signal consist of amplitude (and sometimes phase) information in a narrow frequency range. In order to model the filter-

ing that happens in the cochlea of the human auditory system, gammatone filterbanks (*e.g.,* (Hohmann, 2002; Irino and Patterson, 2006)) is often used as a first step in the process. This section will give a theoretical background of gammatone filterbank which serves as the basis of analysis system used throughout this thesis.

The gammatone filterbank (GF) was first introduced by Patterson et al. (1988) in order to simulate the frequency analysis operated by the human cochlea. Several efficient version of the gammatone filter's implementations were proposed in Cooke (1991); Hohmann (2002) and Holdsworth et al. (1988). For example, an approximation of the 4th-order gammatone filter by a cascade of 1st-order recursive filters was originally proposed by Holdsworth et al. (1988). Cooke (1991) further designed a complex gammatone filter. Hohmann (2002) employed a more detailed implementation of the 4th-order complex gammatone filterbank. The gammatone implementation provided by Hohmann (2002) is used in this thesis.

In particular, Hohmann (2002) used the complex version of the gammatone filter, which has two benefits: (i) the numerator of the z-transform function of the gammatone filter is removed in order to make the implementation simpler; (ii) the envelope of the gammatone impulse response is estimated from the Hilbert envelope of each frequency band signal which is the absolute value of the complex filter output.

According to Hohmann (2002), the impulse response of the complex analog gammatone defined as follow:

$$g_\gamma(t) = t^{\gamma-1}\tilde{a}^t, \quad (t \geq 0) \tag{4.1}$$

$$\text{with} \quad \tilde{a} = \varphi \exp(i\beta)$$

where t is the sample index, γ is the filter order that used to determine the slope of the filter's skirts, φ is the bandwidth parameter which used to determine the duration of the impulse response, and β denotes the oscillation frequency of the filter (phase). Typically, the filter's coefficient $\tilde{a} = \varphi \exp(i\beta)$ is computed from the desired filter bandwidth and centre frequency. The centre frequency of the filter is quantified by the phase, β. Assuming that the desired centre frequency, f_c, and the sampling frequency, f_s, measured in Hz, the numerical value of β written as:

$$\beta = 2\pi \frac{f_c}{f_s}, \tag{4.2}$$

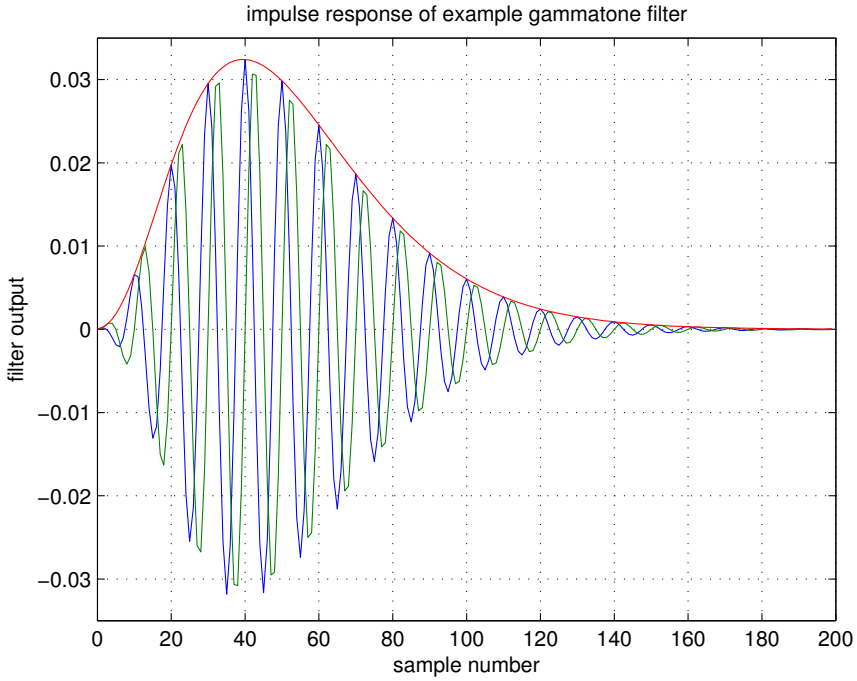

Figure 4.2: The impulse response of an example gammatone filter where the sampling frequency is 10 kHz, centre frequency is 1000 Hz, and 3-dB bandwidth is 100 Hz. The absolute value of the filter output represents the envelope.

Additionally, the frequency response of the Gammatone filter can be expressed as:

$$G_4(z) = \frac{\tilde{a}z^{-1} + 4(\tilde{a}z^{-1})^2 + (\tilde{a}z^{-1})^3}{(1 - \tilde{a}z^{-1})^4} \tag{4.3}$$

which has been simplified by deleting the the numerator of the z-transform in the previous equation since it represents the zeroes in the infinite impulse response (IIR) bandpass filters that result in the following equation:

$$K_4(z) = (G_1(z))^4 = \frac{1}{(1 - \tilde{a}z^{-1})^4} \tag{4.4}$$

The bandwidth of the auditory filter is usually characterised as a function of f_c. It is typically described in either of: (i) -3dB bandwidth or (ii) the Equivalent Rectangular Bandwidth (ERB). The former descriptor, -3dB, defines the differences between the two frequencies when the filter's response has been fallen by a power of two. The ERB of a filter is the bandwidth of a rectangular filter which consists of the same peak gain and passes the same total power for a white noise input (Moore, 1982). The bandwidth of

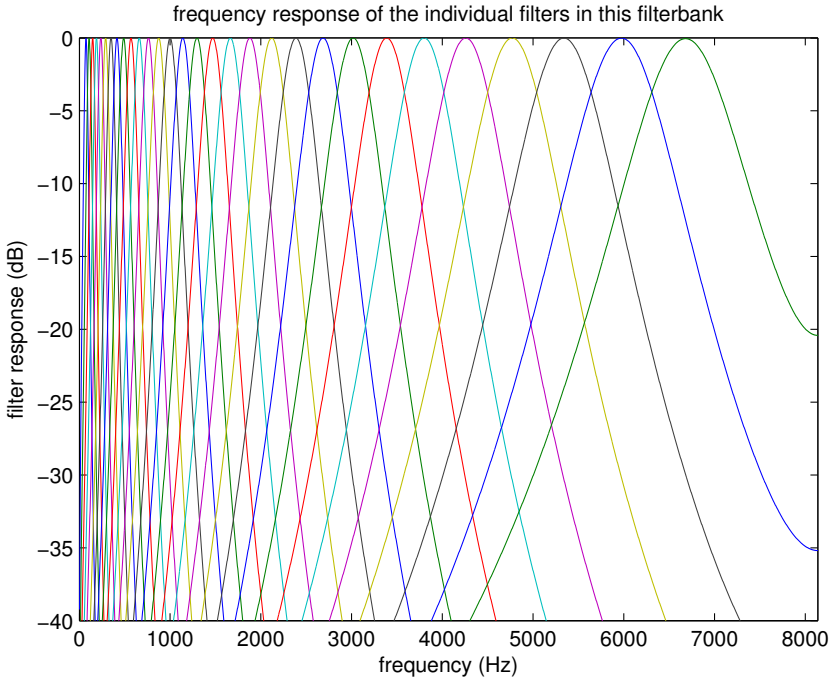

Figure 4.3: Frequency responses of a gammatone filterbank with 30 filters whose centre frequencies are equally spaced between 70 Hz to 6.7 kHz on the ERB-rate scale.

the filters is chosen based on the ERB of human auditory filters. This considered as a measure of the critical bandwidth of human auditory filters (Glasberg and Moore, 1990; Moore and Moore, 2004). Glasberg and Moore (1990) described the accurate match to human data as:

$$ERB_{aud}(f) = l + \frac{f}{q}, \quad \text{with} \quad l = 24.7, q = 9.265 \tag{4.5}$$

The relationship between the bandwidth parameter φ and the ERB of the gammatone filter is investigated by Patterson et al. (1988) and well approximated in the following equation:

$$\varphi = \exp\left(-\frac{2\pi b}{f_s}\right) \tag{4.6}$$

$$\text{with} \quad b = ERB/a_\gamma, \quad \text{and} \quad a_\gamma = \frac{\pi(2\gamma-2)!2^{-(2\gamma-2)}}{(\gamma-1)!^2}$$

where γ is the filter order as in Equation 4.1. Additionally, Patterson et al. (1988) show

that the 3-dB bandwidth f_b in Hz corresponding to a specific ERB is

$$f_b = \frac{c_\gamma}{a_\gamma} ERB \tag{4.7}$$

$$\text{where} \quad c_\gamma = 2\sqrt{2^{1/\gamma} - 1}$$

Furthermore, the centre frequencies for the auditory filterbank are linearly spaced on the ERB frequency scale. The ERB scale is almost logarithmic because of the auditory filters are almost constant Q, which is constant ratio of bandwidth and centre frequency, based on Equation 4.5. The following equation showes the value on the ERB scale as a function of frequency, that is computed by integrating $(\frac{1}{ERB_{aud}(f)})$ across frequency bands:

$$ERB_{scale}(f) = q \cdot \log\left(1 + \frac{f}{l \cdot q}\right) \tag{4.8}$$

$$\Leftrightarrow \quad f = \left(\exp\left(\frac{ERB_{scale}}{q}\right) - 1\right) \cdot l \cdot q, \tag{4.9}$$

$$\text{with} \quad l = 24.7, \quad q = 9.265.$$

Figure 4.2 represents the impulse response of an example gammatone filter where the sampling frequency, f_s, is 10 kHz, centre frequency, f_c, is 1000 Hz, and 3-dB bandwidth, f_b, is 100 Hz. The absolute value of the filter output represents the envelope.

4.2.2 Resynthesis

The speech signal is resynthesised by means of the same parametric representation used in the analysis. The final subsystem of the analysis-resynthesis system, as shown in Figure 4.1, is to resynthesise the audio waveform from the analysis stage's output. Resynthesising the waveform allows the performance of the system to be evaluated through subjective or objective means.

Theoretically, the impulse responses of different frequency bands have different fine structures and group delays. Thus, reconstructing the speech signal is not relatively straightforward and cannot be obtained by simply adding up the filterbank output. This issue can be justified by: (i) the peaks of the impulse responses of the filterbank outputs are not aligned, and (ii) the peak of the envelope of each individual band's impulse response is not aligned with the peak of its fine structure. Evidence from previous research implemented two methods to overcome this problem which are time-reversed

filtering, *e.g.,* (Weintraub, 1985), or phase-correction, *e.g.,* (Holdsworth et al., 1988).

Hohmann (2002) describes a phase-correction method that is based on a peak-alignment in which the fine structure and the envelope of each band's impulse response is delayed, and therefore all bands have their envelope maximum and their fine structure maximum simultaneously. This approach employed throughout this thesis and thus a detailed explanation will be given in this section.

A deficiency in the resynthesis component is introduced if the desired group delay is smaller than the group delay of a particular frequency band. Hohmann (2002) initialised the process of peak-alignment by defining a desired group delay of the system that acts as an alter for the impulse responses to sum up. To start with, the complex filterbank output signals $\tilde{x}_f(t)$ are multiplied with the frequency band-dependent complex factor $\tilde{b}_f$ as follow:

$$\tilde{x'}_f(t) = \tilde{b}_f \tilde{x}_f(t), \quad 0 \leq f < F. \tag{4.10}$$

in which f represents the band index and t represents the sample index and $\tilde{b}_f$ is a phase factor with magnitude 1. When the fine structure is maximum at the (band-dependent) t_f, $\tilde{b}_f$ can be computed as

$$\tilde{b}_f = \exp(i\phi_f) \tag{4.11}$$

$$\text{with} \quad \phi_f = -2\pi f_f t_f \tag{4.12}$$

where f_f denotes the centre frequency of the band f in Hz. Now, for each band, the τ_f is defined according to the following criteria:

case (1) *If the envelope maximum is earlier in time than the desired group delay.*
The real part of the impulse response (fine structure) $x'_f(t)$ has a maximum at the maximum of the envelope. Then, the output is delayed so that envelope maximum and thus the fine structure global maximum meet with the desired group delay. Thus, τ_f is the point in time of the envelope maximum of the respective impulse response.

case (2) *The envelope maximum is later in time than the desired group delay.*
The $x'_f(t)$ has a local maximum at the desired group delay. Thus, τ_f is the desired group delay.

After that, the real parts $x'_f(t)$ are delayed by a band-dependent amount of Δt_f samples as:

$$x''_f(t) = x'_f(t - \Delta t_f) \tag{4.13}$$

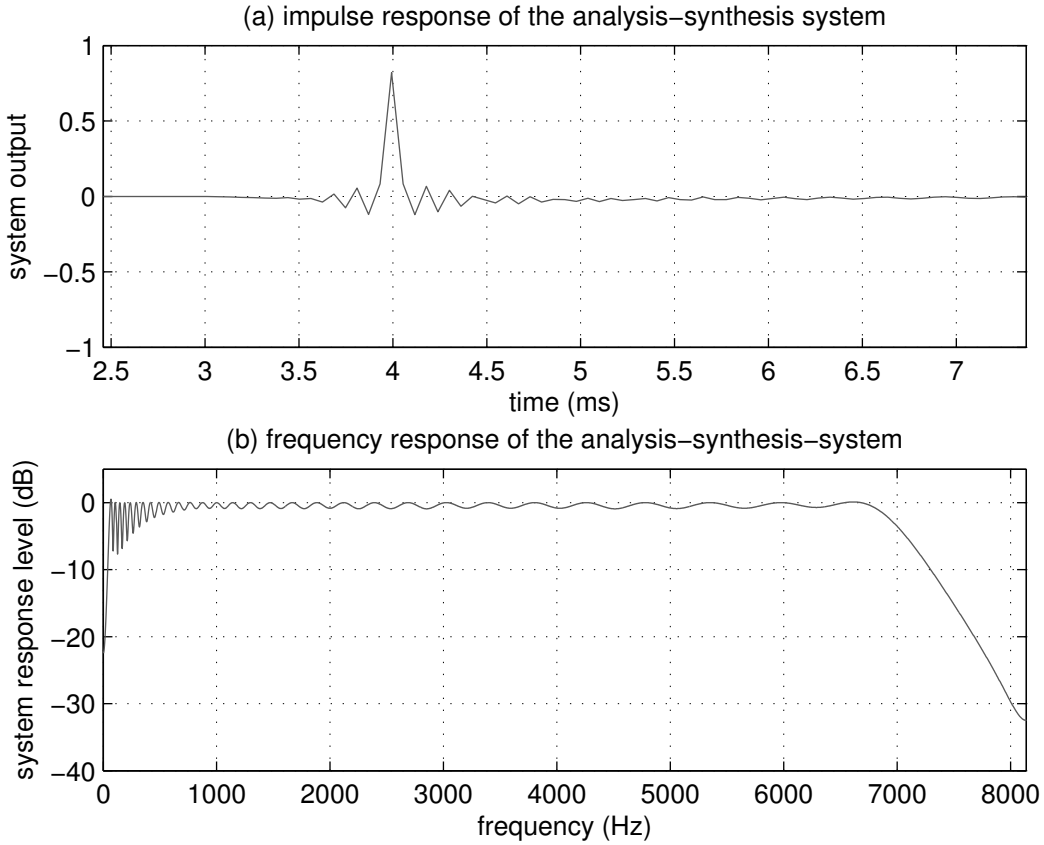

Figure 4.4: The impulse response of the analysis-resynthesis system using a gammatone filterbank (a peaked impulse response is obtained at the desired group delay of 4 ms) (in (a)), and magnitude of the transfer function of the analysis-resynthesis system using the gammatone filterbank (in (b)).

$$\text{where} \quad \Delta t_f = \text{int}\left(\Delta\tau_f f_s\right) \tag{4.14}$$

in which, int(), denotes the nearest integer operation. In case (1) the τ_f denotes the differences in time between the desired group delay and the point in time of the envelope maximum of the respective impulse response, whereas in case (2) $\tau_f = 0$.

Finally, the resynthesised speech signal is obtained by a weighted sum across all frequency bands using Equation 4.13 and 4.14 as:

$$x'''(t) = \sum_{f=0}^{F-1} g_f x''_f(t) \tag{4.15}$$

where g_f is the band-dependent weights.

For a demonstration, Figure 4.3 shows the magnitude frequency response of 30 gammatone filters that are equally distributed along the ERB scale from 70 Hz to 6.7 kHz at at f_s 16276 Hz. The filter bandwidth (ERB) was defined to be one ERB_{aud}. Figure 4.4 displays the impulse response that adapted the resynthesis procedure described in Section 4.2.2 (case 2). It was implemented to the example filterbank design using a desired group delay of 4 ms (65 samples at f_s 16276 Hz). Because of the incomplete compensation of the group delays at the lower frequency bands, the impulse response is peaked with some small side maxima (where case 2 can be applicable in this scenario). Therefore, there is some some minor ripples in the magnitude response and the group delay is larger than the desired value.

4.3 Signal Modification

Having a representation of signal features creates many opportunities for modifications of the signal prior to and/or after resynthesis stage. The aim of this section is to highlight possible modification strategies. Kleijn et al. (2015) categorise several modification strategies some of which have been described in Chapter 3. Examples categories might include: temporal scale, spectral scale, spectro-temporal scale, cepstral scale modifications.

Temporal scale approach modifies the temporal evolution of the speech signal without a corresponding change in the spectral content. This can be accomplished by overlapping and adding small windowed fragments of the waveform or multiplying all or some of the temporal information by a specified amount. It is applied for compressing or stretching speech. Furthermore, it can be made to manipulate within amplitude scale as well in which the speech signal can be amplified or attenuated.

Spectral-scale approach, on the other hand, modifies the spectral characteristic while retaining the amplitude and temporal content of the signal. It works by multiplying all or some of the spectral information by a constant amount (Maher, 1991).

Spectro-temporal scale approach refers to a change in both time and frequency features of speech. The modification in this scale is typically obtained by multiplying all or some of the target feature by a gain. The computation of the gain differs according to the application. Having a view of and control over both time and frequency aspects of the speech promotes a modification in a cepstral-domain.

Further possible modifications used in the literature can be categorised based on a number of different criteria. Perhaps the simplest classification can be made based

on its dependency on the environmental noise. For this case, the modification can be noise-dependent or -independent. Further general class of modification is characterised as time-varying or time-invariant, and as linear or nonlinear (*e.g.,* Taal et al. (2013) use a linear filtering of the signal).

Addition classification criteria presented in the literature according to Kleijn et al. (2015) can be made according to its the level of processing implemented on the spoken message which can be categorised into: (i) lexical, (ii) prosodic, and (iii) spectral and temporal modifications and might depend on the environmental noise, too. Finally, taking into account the origin of a modification, there is a class based on mimicking strategies maintained consciously or subconsciously by humans in adverse listening conditions. Example modifications include pitch modification, vowel space adjustment, and uniform speaking rate reduction (Cooke et al., 2014, 2013a). Further, mimicking rational strategies that is based on experts insights in the human auditory periphery and in cognition has been investigated by Niederjohn and Grotelueschen (1976).

Moreover, the *spectral and temporal* modifications can be subdivided into spectral, temporal, and spatial signal modifications. For instance, a straightforward spectral shaping is employed by Sauert and Vary (2009); Taal et al. (2013) and Zorila et al. (2012). An example of spectro-temporal energy redistribution is considered in Tang and Cooke (2012) where the glimpse proportion is optimised, and a low-complexity approach using spectro-temporal energy redistribution by optimising a perceptual distortion measure is presented in Taal et al. (2014).

4.4 Optimising Intelligibility

As mentioned earlier, the principle of the closed-loop near-end intelligibility enhancement discussed in this thesis is to select an intelligibility measure and a modification strategy, and then adjust the parameters of the modification strategy in order to maximise the measure. This section defines the optimisation problem of intelligibility enhancement system in Section 4.4.1, and outlines possible optimisation classes that can be used in Section 4.4.2. Finally, in Section 4.4.3, we give a detailed description of the chosen optimisation method.

4.4.1 General Problem Formulation

The near-end intelligibility enhancement problem is typically posed as follows; given a set of δ parameters and a noise signal, n, one could derive a new set of δ that maximises a chosen intelligibility measure as shown in Figure 4.1. The optimisation problem can then be written as:

$$\tilde{\delta} = \underset{\delta}{\operatorname{argmax}}\, \mathcal{J}(\delta|x, n), \tag{4.16}$$

where x denotes the clean speech signal and $\mathcal{J}(.)$ is the objective function to be maximised. For simplification, we refer $\mathcal{J}(\delta)$ to the objective function in Equation 4.16.

4.4.2 Solving the Optimisation Problem

Generally, a mathematical problem of either minimising or maximising an objective function can be solved depending on whether the optimisation problem is: (i) continuous or discrete, (ii) constrained or unconstrained, (iii) and finally, local or global. It is arguable that the most important distinction arising as the results of having constraints on the input parameters or not (Nocedal and Wright, 2006). This section gives an general overview of the differences between these solutions, readers are referred to Nocedal and Wright (2006) for more information about numerical optimisation methods.

First, discrete optimisation techniques are usually implemented if one attempts to find a solution which is one of a number of objects in a finite set. Continuous optimisation techniques, on the other hand, is applied to find a solution from an uncountable infinite (*i.e.,* a set of vectors with real components). It is claimed that continuous optimisation problems are easy to solve due to the smoothness of the objective functions (*i.e.,* differentiable). This makes it possible to use objective and constraint information at a specified point to deduce information in regard to the function's behaviour at all points nearby.

In addition, the optimisation problem can be solved regarding the nature of the objective function and constraints (linear, nonlinear, convex). In particular, unconstrained optimisation problems are usually used if there is not explicit constraints on the parameters. By contrast, constrained problems implemented when there is constraints that are either simple bounds, a more general linear constrains, or nonlinear inequalities between the parameters.

Fastest optimisation algorithms are usually attempting to find a local solution rather than searching for the best of all such minima, known as a global solution. The latter is

typically hard to identify and locate. An example algorithms are convex optimisation, whereby all local solutions are treated to be also global solutions. Nevertheless, in the nonlinear optimisation including constrained and unconstrained, the local solutions are not global solutions.

Based on that, the method that will be adapted for solving the optimisation problems in this thesis should meet the following: (i) robustness, (ii) efficiency, and (iii) accuracy. First, the method should perform well for all reasonable choices of the initial setting of parameters. Second, it should not be computationally-expensive. Finally, it should be able to identify a solution with precision. In the following section, we will describe Nelder-Mead simplex optimisation.

4.4.3 Nelder-Mead Simplex Optimisation

The Nelder-Mead (NM) method, developed by Nelder and Mead (1965), works by maximising a scalar-valued nonlinear function of a number of real variables using only function values, without any derivative information. This method is based on the iterative update of a simplex.

As the function, $\mathcal{J}(\delta)$, we are maximising with respect to δ parameters might be non-convex and not necessarily have derivatives, we use a direct NM simplex search method. The NM simples search method designed for solving the unconstrained optimisation problem. The method designed for finding a local maximum of a function that has several variables (Lagarias et al., 1998; Nelder and Mead, 1965).

Let $\mathcal{J}$ be our objective function of m dimension. A simplex is a geometric shape in m dimensions that is the convex of $m+1$ vertices. The simplex is represented by vertices $\delta_1, \delta_2, \cdots, \delta_{m+1}$ by Δ. The NM method iteratively yields a sequence of simplexes to approximate an optimal solution $\tilde{\delta}$ of Equation 4.16. In each individual iteration k, the vertices $\{\delta_i, i = 1, \cdots, m+1\}$ of the simplex are sorted using the objective function values as:

$$\mathcal{J}(\delta_1) \leq \mathcal{J}(\delta_2) \leq \cdots \leq \mathcal{J}(\delta_{m+1}). \tag{4.17}$$

where δ_1 denotes the best vertex, and δ_{m+1} denotes the worst.

There are four alternative operations used in the NM method: *reflection*, *expansion*, *contraction*, and *shrink*, each of which is associated with a scalar parameter: ρ_1 (reflection), ρ_2 (expansion), ρ_3 (contraction), and ρ_4 (shrink). The values of these parameters

have the following conditions:

$$\rho_1 > 0, \quad \rho_2 > 1, \quad 0 < \rho_3 < 1, \quad \text{and} \quad 0 < \rho_4 < 1.$$

According to the standard implementation of the NM method, the parameters are defined to be

$$\rho_1 = 1, \quad \rho_2 = 2, \quad \rho_3 = \frac{1}{2}, \quad \text{and} \quad \rho_4 = \frac{1}{2}.$$

Furthermore, we assume that $\bar{\delta}$ be the centroid of the m best vertices which is defined as

$$\bar{\delta} = \frac{1}{m} \sum_{i=1}^{m} \delta_i$$

A summary of one iteration of the Nelder-Mead (NM) algorithm, as described by Lagarias et al. (1998), can be presented as follows:

Summary of one iteration of the Nelder-Mead (NM) algorithm

1. **Order.** Evaluate $\mathcal{J}$ at the $m+1$ vertices of Δ and order the vertices so that Equation 4.17 implemented.

2. **Reflection.**

   (a) Calculate the reflection point δ_r as follow

   $$\delta_r = \bar{\delta} + \rho_1(\bar{\delta} - \delta_{m+1})$$

   (b) Evaluate $\mathcal{J}_r = \mathcal{J}(\delta_r)$. If $\mathcal{J}_1 \leq \mathcal{J}_r < \mathcal{J}_m$, replace δ_{m+1} with δ_r.

3. **Expansion**.

   (a) If $\mathcal{J}_r < \mathcal{J}_1$ then calculate the expansion point δ_e. as follow

   $$\delta_e = \bar{\delta} + \rho_2(\delta_r - \bar{\delta})$$

   (b) Assess $\mathcal{J}_e = \mathcal{J}(\delta_e)$. If $\mathcal{J}_e < \mathcal{J}_r$, replace δ_{m+1} with δ_e; otherwise replace δ_{m+1} with δ_r.

4. **Outside Contraction.**

   (a) If $\mathcal{J}_m \leq \mathcal{J}_r < \mathcal{J}_{m+1}$, calculate the outside contraction point δ_{oc} as

   $$\delta_{oc} = \bar{\delta} + \rho_3(\delta_r - \bar{\delta})$$

   (b) Assess $\mathcal{J}_{oc} = \mathcal{J}(\delta_{oc})$. If $\mathcal{J}_{oc} \leq \mathcal{J}_r$, replace δ_{m+1} with δ_{oc}; if the condition does not apply go to step 6.

5. **Inside Contraction.**

   (a) If $\mathcal{J}_r \geq \mathcal{J}_{m+1}$, calculate the inside contraction point δ_{ic} as follow

   $$\delta_{ic} = \bar{\delta} - \rho_3(\delta_r - \bar{\delta})$$

   (b) Evaluate $\mathcal{J}_{ic} = \mathcal{J}(\delta_{ic})$. If $\mathcal{J}_{ic} < \mathcal{J}_{m+1}$, replace δ_{m+1} with δ_{ic}; if the condition does not apply go to step 6.

6. **Shrink.** For $2 \leq i \leq m+1$, define

   $$\delta_i = \delta_1 - \rho_4(\delta_i - \delta_1)$$

4.5 Chapter Summary

This chapter is a core chapter in this thesis, and has been devoted to a closed-loop optimisation based near-end intelligibility enhancement framework by means of gammatone filterbank analysis-resynthesis. First, it introduced basic notions of speech analysis and provided an in depth explanation of gammatone analysis system. Next, we examined the possible classes of speech manipulation with the aim of optimising the intelligibility. An efficient way to resynthesis speech was then presented using gammatone filterbank and peak-alignment method. The final section of this chapter dealt with solving the optimisation problem that employ unconstrained optimisation method known as Nelder-Mead simplex direct search method.

In the following three chapters, we will see how can this framework be implemented by selecting an intelligibility model and a modification strategy. Starting with Chapter 5, we will use a measure of energetic masking and a spectral modification strategy.

Chapter 5

Spectral Modification Based on Glimpse Proportion Measure

5.1 Introduction

Pre-enhancement generally works by making the speech harder to mask. In Lombard speech this equates to increasing the intensity, increasing f_0, lengthening vowel duration and reducing spectral tilt to boost the high frequencies (Junqua, 1993; Lu and Cooke, 2008; Van Summers et al., 1988), readers are referred to Section 3.2 for more information about Lombard speech. Noise-adaptive algorithms in particular make similar changes although generally a fixed-intensity constraint is applied because it is undesirable to maintain intelligibility by simply boosting the signal energy, readers are referred to Section 3.3.2 for a review about noise-adaptive algorithms.

As mentioned in the previous chapter, the challenge for pre-enhancement systems is to *optimally* adjust the parameters of the speech modification algorithm. Typically the parameters may be tuned by using knowledge of the background noise and an objective intelligibility model, *i.e.,* they are adjusted so as to maximise the predicted intelligibility. This automated closed-loop approach allows the adoption of highly flexible near-end enhancement algorithms that can finely control the acoustic features of the speech. We thus define, in the previous chapter, a general framework for a closed-loop near-end intelligibility enhancement system. In this chapter, we implement this framework using an intelligibility model and a speech modification strategy.

The starting point of implementing the closed-loop near-end intelligibility enhancement framework is to select a model of speech intelligibility. The choice here is made based on an auditory masking model used in the missing data theory (*e.g.,* (Cooke et al.,

2001)). The missing data theory is based on two assumption: First, it is possible to estimate prior to decoding which spectro-temporal (S-T) elements belong to speech and which belong to background noise that occludes the speech which turn into a binary representation using the auditory masking model. Second, the binary representation can then be used to guide the decoder to ignore these elements, or to replace the occluded elements by clean speech estimates prior to decoding. The intelligibility measure is defined therefore from the auditory masking model (sometimes known as a glimpsing model) using the derived binary representation in order to minimise energetic masking. This measure is known as Glimpse Proportion (GP) (Cooke, 2006). A more detailed description of how the measure can be used to make a prediction of intelligibility will be given later.

The second is to apply a modification strategy that can be fitted within the closed-loop near-end intelligibility enhancement framework. The modification attempts to mimic to some extent the the spectrum of speech produced in noise in order to make the speech harder to interfere. A more detailed explanation is provided later on of how the modification works.

This chapter is organised into four sections. In Sections 5.2, we provide the essential background about the Glimpses Proportion measure. In Section 5.3, we describe a spectral modification strategy using cepstral coefficient analysis. Section 5.4 shows the derivation of an optimisation problem using the spectral modification strategy and the Glimpses Proportion measure. In Section 5.5 and 5.6, the objective and subjective evaluation of the near-end intelligibility enhancement system are presented. Finally, the chapter concludes with a general discussion in Section 5.7 and a summary in Section 5.8.

5.2 The Glimpses Proportion Measure

The Glimpses Proportion (GP) measure is derived from the Glimpse model for speech perception in noise reported in Cooke (2003). The model was motivated by the idea that the S-T overlap between speech signal and background noise is a main cause of energetic masking. The model therefore identifies elements of relatively higher SNR wherein speech information which remains unmasked, labelled as 'glimpses'. The definition of 'glimpses' used in this thesis is similar to that used in Cooke (2006).

The decision metric employed by the GP, mainly considers the audibility of the speech in noise, quantified by the number of identified glimpses of a given speech signal in a given masker. Research has shown that the GP is a good predictor of intelligibility

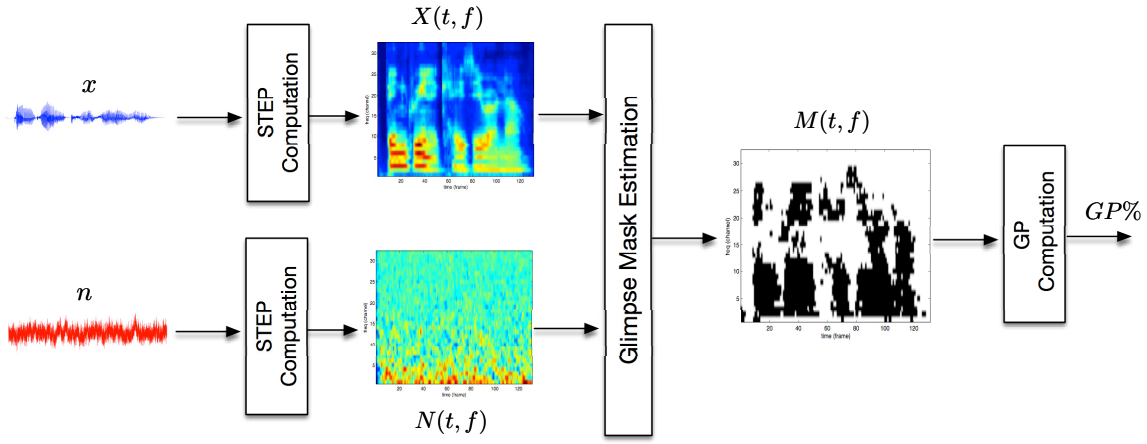

Figure 5.1: Block diagram of the overall structure of the Glimpse Proportion (GP) measure, where x represents a speech signal, n is the environmental noise, $X(t,f)$ and $N(t,f)$ are the Spectro-Temporal Excitation Pattern (STEP) of speech and noise respectively. To obtain the STEP representation, the desired signal is first processed by a bank of 32 gammatone filters equally spaced between between 50 and 8000 Hz. For each spectral band, the Hilbert envelope of the filter bank is computed and smoothed with a leaky integrator with an 8 ms time constant. The smoothed envelope is then downsampled to 100 Hz and log-compressed. The glimpses mask, $M(t,f)$, is calculated using the $X(t,f)$ and $N(t,f)$ as a spectro-temporal (S-T) region wherein the speech power is greater than the noise power limited by some threshold value. The overall number of identified glimpses is converted into a percentage to form the GP.

because it correlates well with subjective scores for intelligibility of natural speech in noise (Cooke, 2006).

Figure 5.1 illustrates a sketch of the overall structure of the GP measure considered in this thesis. To compute the GP, the glimpses mask need to be estimated. To do that, first speech x and noise n signals are converted into Spectro-Temporal Excitation Pattern (STEP) representation. To obtain the STEP representation, a number of steps have to be taken. First, the signal is filtered using a bank of F gammatone filters with centre frequencies, f_c, spread evenly on an the Equivalent Rectangular Bandwidth (ERB) scale (Moore and Glasberg, 1996) with filter bandwidths, λ, matched to the ERB of human auditory filters. The gammatone filterbank used here is based on the work of Hohmann (2002) (for more details see Section 4.2.1 in Chapter 4). The instantaneous Hilbert envelope of each gammatone filter output is then computed with an absolute value operation. This envelope is smoothed with a leaky integrator with an τ ms time constant using a first-order lowpass filter. The impulse response of the first-order lowpass

filter is defined by the parameter τ as follow:

$$J_1(z) = \frac{1 - \exp(-\alpha)}{1 - \exp(-\alpha)\, z^{-1}}, \quad \alpha = \frac{1}{\tau f_s}. \tag{5.1}$$

The smoothed envelope is then decimated by a factor, L. By taking the logarithm of the down sampled spectrum, $\tilde{X}(t, f)$, the results forms the STEP representation, $X(t, f)$, where $X(t, f) = \log(\tilde{X}(t, f))$, that is used as the basis of the intelligibility models described here.

Following the computation of STEP representation of speech, $X(t, f)$, and noise, $N(t, f)$, the glimpse mask , $M(t, f)$ is computed. It is defined as a binary matrix within which 1 denotes that the speech energy in the corresponding S-T element exceeds the masker energy by a predefined threshold, θ, and 0 denotes otherwise. The threshold is called the local SNR criterion measured in dB. Precisely, the mask can thereafter be expressed as;

$$M(t, f) = \begin{cases} 1, & \text{if } x(t, f) > n(t, f) + \theta \\ 0, & \text{otherwise} \end{cases} \tag{5.2}$$

where $x(t, f)$ denotes the speech energy within the element of time frame t and frequency channel f and $n(t, f)$ is the masker energy in the S-T element. The mask calculation requires access to the speech, x, and noise, n, signals prior to computation. The GP measure is then computed as the percentage of the overall number of identified glimpses. The GP can hence be expressed as follow;

$$GP(x, n) = \frac{100}{TF} \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} \mathcal{H}(M(t, f)). \tag{5.3}$$

where T and F denote the numbers of time frames and frequency bands. $\mathcal{H}(.)$ is the Heaviside step function counting the number of 'glimpses' which meet the local audibility criterion θ.

5.3 Spectral Modification Based on Cepstral Coefficients

The previous section gave a detailed explanation of the GP measure. This section describes a spectral modification method based on cepstral analysis that can be employed within the closed-loop enhancement framework defined in Chapter 4.

Generally speaking, the pre-enhancement techniques mimic strategies that talkers themselves adopt to counteract the effects of background additive noise (*e.g.,* the Lombard effect (Lombard, 1911)). The acoustic changes in Lombard speech includes an increase in intensity, reduction in speaking rate, increase in f_0, increase in vowel duration, a shift in the energy distribution of the spectral content from low to middle and high frequency regions which results in flatter spectral tilt, increase in the first formant and in some studies increases in the second formant were also observed (see Section 3.2 for more details about acoustic changes in Lombard speech). Spectral shaping techniques has also been developed in many near-end intelligibility enhancement systems due to its potential to provide significant intelligibility gains with low computational complexity (Cooke et al., 2013a).

In particular, for speech produced in noise (*i.e.,* Lombard speech), a talker modifies the tilt of the spectrum by reducing spectral tilt to boost the high frequencies in order to improve the intelligibility in noise (*e.g.,* (Junqua, 1993; Lu and Cooke, 2008; Van Summers et al., 1988)). Thus, it is hypothesised that modifying the spectral shape of speech to hopefully match that in Lombard speech might improve the intelligibility of speech. To do so a spectral shape of a spectral envelope need to be computed. In this work, the spectral shape is defined using parameters of the cepstral analysis in the erb-scale.

An arbitrary reshaping could be represented as F independent scaling factors, $S_c(f)$, where f represents the frequency band index, *i.e.,* in the log domain using cepstral coefficients c_n. The c_n appear to represent a more efficient representation of speech spectra than other analysis methods as reported in Deng and O'Shaughnessy (2003). Thus, for a particular frequency band, f, the spectral shaping can be derived as:

$$S_c(f) = \sum_{n=0}^{N-1} c_n \cos(\frac{\pi}{F}(n + \frac{1}{2})f), \quad f = 1, \ldots, F. \tag{5.4}$$

The first value c_0 indicates the average speech power. The following parameters c_1 represents the balance of power between low and high frequencies. For each $n > 1$, c_n represents increasingly finer spectral detail (as a cosine with n periods weights smaller frequency ranges with its alternating oscillations) (Deng and O'Shaughnessy, 2003).

The spectrum is then shaped by applying a band-dependent scaling to the $X(t, f)$, before re-summing them to form the enhanced signal:

$$\hat{X}_c(t, f) = \sum_{f=1}^{F} (X(t, f) + S_c(f)) \tag{5.5}$$

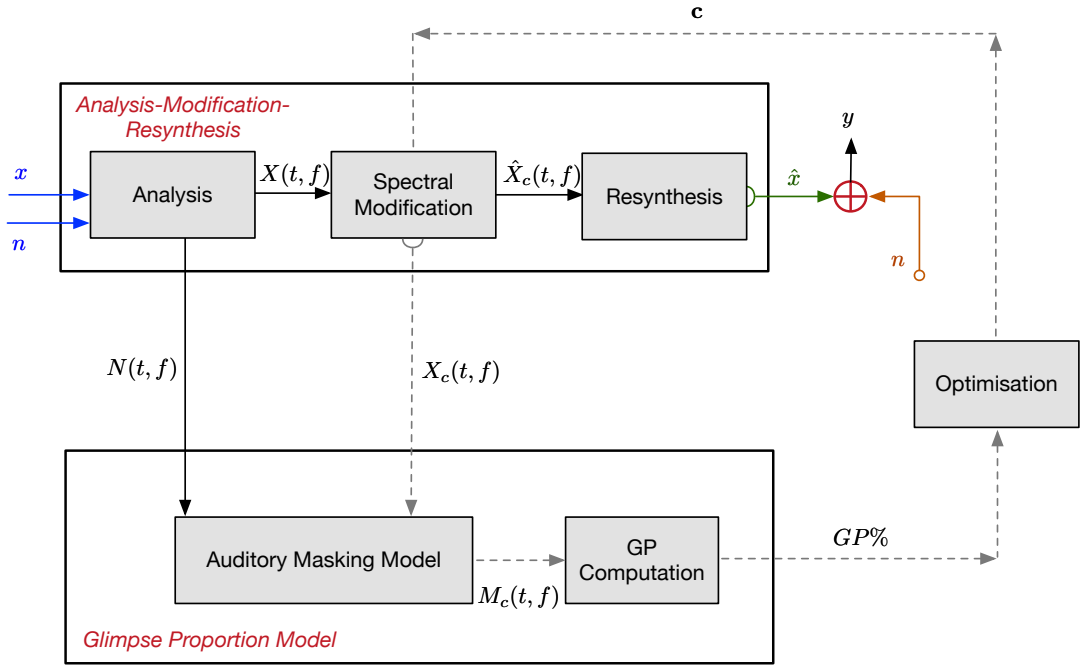

Figure 5.2: A schematic diagram of optimisation process for the GP-based spectral modification system, where x and n denote the input speech and noise signals. The $X(t, f)$ and $N(t, f)$ are the Spectro-Temporal Excitation Pattern (STEP) of x and n, respectively. The $X_c(t, f)$ denotes the modified and re-normalised STEP, used to compute the glimpses mask $M_c(t, f)$. The $\mathbf{c}$ represents the optimal cepstral coefficients and GP_c is the Glimpse proportion metric. Finally, the optimal setting will result in the optimal STEP of speech $\hat{X}_c(t, f)$ in which it is resynthesised to represent the optimal enhanced and re-normalised signal $\hat{x}$. Solid arrows indicate fixed input to the optimisation process, whereas the grey dashed arrows indicate iterative process of optimisation.

Note, using this formulation the F spectral shaping weights are controlled by N parameters, $c_0, \ldots, c_N$. Further, c_0 is arbitrarily fixed to 0 because it simply adds a constant gain factor across frequency that does not change the spectral shape.

5.4 Optimising Intelligibility

In this section, the optimisation process for the GP-based spectral modification system taken to obtain the optimal cepstral coefficients is described (as illustrated in Figure 5.2). It works by optimising the parameters of the adaptive spectral modification (described in section 5.3) using the GP measure (described in section 5.2). It is assumed that the

speech and noise signals are known and the parameters of the spectral-shaper are optimised to minimise the perceptual masking of the speech signal under energy preservation constraint.

In principle, it is possible to *pre-shape* the spectrum of the input speech signal to adjust the gain of band-energies of the output speech by reducing the degree of masking through maximising the GP. The modification to the original spectrum is tuneable to c_n over the duration of utterance.

The near-end intelligibility enhancement can now be optimised by searching for the optimal cepstral parameters $\hat{\mathbf{c}}$, where $\hat{\mathbf{c}} = [\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_n]^T$, that maximise the GP measure for a given speech x and noise n, *i.e.,* the optimal parameter values are given by,

$$\hat{\mathbf{c}} = \underset{\mathbf{c}}{\operatorname{argmax}}\, GP\left(x, n; c\right), \tag{5.6}$$

$$\text{where} \quad GP\left(x, n; c\right) = \frac{100}{TF} \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} \mathcal{H}\left(\hat{X}_c(t,f) > \left(N(t,f) + \theta\right)\right). \tag{5.7}$$

in which T and F denote the numbers of time frames and frequency bands. The $\hat{X}_c(t,f)$ and $N(t,f)$ denote the STEP of the pre-enhanced speech and noise at time frame t and frequency band f, respectively. $\mathcal{H}\left(.\right)$ is the Heaviside step function counting the number of 'glimpses' which meet the local audibility criterion θ. The optimisation problem has been solved using the Nelder-Mead Direct Search method (described in Chapter 4 (Section 4.4.3).

5.5 Objective Evaluation

The general aim of GP-based spectral modification system is to manipulate speech at source to improve the intelligibility of speech in noise while keeping the overall energy level unchanged. In this section, the GP-based spectral modification system has been evaluated by means of acoustic analysis and macroscopic intelligibility measures. For acoustic analysis, we used relative average spectra (*i.e.,* spectral shape). In macroscopic intelligibility evaluation, the GP (described in Section 5.2) and Short-Term Objective Intelligibility measure (STOI) proposed by Taal et al. (2011) has been used.

It should be noted that the GP measure, to be used in the objective evaluation, has been integrated into the optimisation cost function to guide the design of the system. Thus, in addition to the GP measure we will use the STOI measure, which has shown high correlations with speech intelligibility in Taal et al. (2014). The STOI measure

is based on the correlation between a reference undistorted speech signal and a distorted speech signal. In particular, the measure is a numeric estimate computed as a short-term correlation coefficient between the clean reference speech envelope and the distorted speech envelope at the output of a number of 1/3-octave bandpass filters. An important process in the STOI model is the normalisation and clipping of the processed envelope in order to remove the effects of level differences between the two signals from the correlation coefficient. As a result, STOI effectively measures the similarity of the modulation content from the envelope waveforms of the two signals, whereby any reduction of the correlation may be assumed to result from noise modulations or other non-speech modulations. The model computed as an index between 0 and 1, in which 0 represents maximum distortion and 1 represents minimum distortion of the speech signal (Taal et al., 2011).

This section is organised into five parts: The corpus used in the evaluation and the experimental setup are described in Section 5.5.1. In Section 5.5.2, we investigate the role of using different number of cepstral coefficients on intelligibility improvement. In Section 5.5.3, we study the behaviour of the system in stationary and non-stationary maskers. Section 5.5.4 answers the question of whether males and females differs in their intelligibility benefit from the enhancement system. Finally, in Section 5.5.5, we study how the optimised speech, *i.e.,* the output of the system that has been optimised for stationary noise, behaves in different instances of white noise.

5.5.1 Grid Corpus and Experimental Setup

This section presents the corpus and experimental setup employed in this thesis. Where appropriate, in each chapter the experimental setup may be briefly repeated to make the thesis easier to understand. Speech recogniser setup related to each experiment will be given at the beginning of each chapter.

Grid Corpus

The speech materials are from Grid corpus (Cooke et al., 2006). The corpus consists of sentences recorded by a total of 34 native English speakers (18 male and 16 female). All sentences exhibit the same six words with a fixed grammar of the form <*command*><*colour*> <*preposition*> <*letter*> <*number*> <*adverb*>, as shown in Table 5.1 (*e.g.,* 'bin green at k z now'). There are 1000 utterances recorded from each speaker sampled at 25 kHz, producing a total of 34,000 sentences. The corpus has four colour

Table 5.1: Structures of the sentences in the Grid corpus.

Verb	*Colour*	*Prep.*	*Letter*	*Digit*	*Adverb*
bin	blue	at	a-z	1-9	again
lay	green	by	(no 'w')	and zero	now
place	red	in			please
set	white	with			soon

choices which are 'red', 'green', 'blue' and 'white', while 25 letters are from the English alphabet ('w' was removed due to its multi-syllabicity) and finally the 10 digits are from '0' to '9'. The length of each utterance is about 2.2 seconds. A test set consists of 300 utterances (around 150 utterances males and the same for females talker) were randomly drawn from the Grid corpus.

Since the aim of this thesis is to use models of speech intelligibility during the development phase of the pre-enhancement system, it is crucial to choose a corpus that can accommodate with the nature of the models *i.e.,* whether it is macroscopic or microscopic. Grid corpus provides us with controlled and phonetically balanced speech materials that comprises tokens whose durations make them applicable to use with speech perception studies as well as behavioural studies. It also has large amounts of data for each speaker, and thus making it applicable for training statistical speech models (Barker and Cooke, 2007; Cooke et al., 2006).

Noise materials

As additive disturbance two different noises were considered: (i) a stationary speech-shaped noise (SSN) that was generated by filtering white Gaussian noise through a 100-order all-pole filter, the long-term average spectrum of this noise was approximated to match that of the Grid speech material, and (ii) a non-stationary **N**-talker babble modulated noise (BMN), which were produced by modulating SSN with the envelope of **N**-talker babble for various **N**. As in Cooke (2006), the envelope was calculated by convolving the absolute value of an **N**-talker babble signal with a 7.2 ms rectangular window. Babble was generated then by summing utterances with equal rms energy from the Grid corpus. In this study, **N** was set to 5.

Speech types

We consider three speech types for evaluation in this chapter two of which are modified speech and the original (unmodified) speech 'ORG', as illustrated in Table 5.2. The modified speech types include the GP-optimised speech 'GP-OPT' and a reference system 'SS' proposed by Zorila et al. (2012), in which only the first part of their system (*i.e.,* spectral shaping) is used. The SS is used later on in the listening test which will be described in the next section.

Table 5.2: Speech types used for the evaluation in this chapter.

Type	*System*	*Required Knowledge*
ORG	Original unmodified speech	-
GP-OPT	GP-optimised speech using spectral shaping modification technique (described in Section 5.3)	Speech and noise signals
SS	modified speech using spectral shaping developed by Zorila et al. (2012) (*i.e.,* only the first part of their system is implemented here)	Speech signal

To generate the GP-OPT modified speech, the optimal cepstral parameters are first estimated using the optimisation [1] described in Section 5.4. Then the spectrum of speech shaped in the analysis-modification-resynthesis framework. In particular, the speech signal is filtered using a bank of 32 gammatone filters with centre frequencies spread evenly on the ERB scale between 50 and 8000 Hz with filter bandwidths matched to the ERB of human auditory filters. After that, the envelope of each gammatone filter output is computed. This envelope is then smoothed by a first-order low-pass filter with an 8 ms time constant. Then, the smoothed envelope is down-sampled to 100 Hz. Following down-sampling, the amplitude envelope is logged to turn the amplitude into the log-energy domain. The spectrum is then shaped by applying a band-dependent scaling using the optimal cepstral parameters to the gammatone filter outputs before re-summing them to form the *pre-enhanced* signal, $\hat{x}$. Re-synthesising is then applied to generate the spectrally shaped speech signal. Care needs to be taken when summing the bands to

[1] The optimisation performed using fminsearch algorithm in matlab. The algorithm allow to define a particular termination criteria, according to both the absolute size of the simplex, the difference between the highest and the lowest function value in the simplex, and the number of iterations.

compensate for band-dependent phase delays introduced by the analysis (for details see Section 4.2.2). After re-synthesising the spectrally shaped signal is re-normalised such that the global signal energy remains unchanged before and after spectral manipulation. The result $\hat{x}$ signal was mixed with the masker at a desired SNR level. Note that the number of iteration used to obtain the optimal cepstral coefficients was 25 iterations.

5.5.2 Performance analysis using different number of cepstral coefficients

The aim of this analysis is to investigate the impact of using different number of cepstral coefficients on the shape of the spectra in a steady-state masker, SSN. The 'optimal spectral shape' (sometimes we refer to it as relative spectra) is computed using the optimal setting of cepstral coefficients after the optimisation process. In particular, the optimal spectral shape is computed as the log differences between the average spectral envelopes calculated using all frames of the ORG and modified speech. The resulting optimal spectral shape are then averaged and represented in (dB).

Figure 5.3 shows the spectral gain in (dB) of GP-OPT modified speech for SSN using different number of cepstral coefficients, N, at a sentence level and averaged across the test set. For instance, the spectral gain in case (a) where $N = 0$ represents a flat tilt or no shaping which is the case of original unmodified speech, the spectral shape in case (b) where $N = 1$ is generated using the first cepstral coefficient (c_1), and the spectral shape in case (c) where $N = 2$ is generated using the first and second cepstral coefficients (c_1 and c_2). One interpretation of these results is that as the number of cepstral coefficients N increases, more spectral details is introduced (as shown in case (f) in which $N = 10$ compared to case (a) in which $N = 1$ in Figure 5.3). Therefore, the time in which the optimisation takes to converge to the optimal solution is increased as N increases.

Figure 5.4 shows an example illustration of the glimpse mask using different number of of cepstral coefficients N, varies from 0 to 10. The mask was generated for speech utterance of a male talker "bin green at k z now" and SSN with equal overall rms levels and at a threshold, θ, of 0 dB, where white pixels indicate 0 and black pixels indicate 1. The mask in case (a) where $N = 0$ is generated using the original unshaped speech, the mask in case (b) where $N = 1$ is generated using the modified speech whose spectrum is tilted by the first cepstral coefficient (c_1). Furthermore, the mask in case (c) where $N = 2$ is generated by the modified speech whose spectrum is shaped by the first and second cepstral coefficients (c_1, and c_2), and so on. It can be seen that the amount of

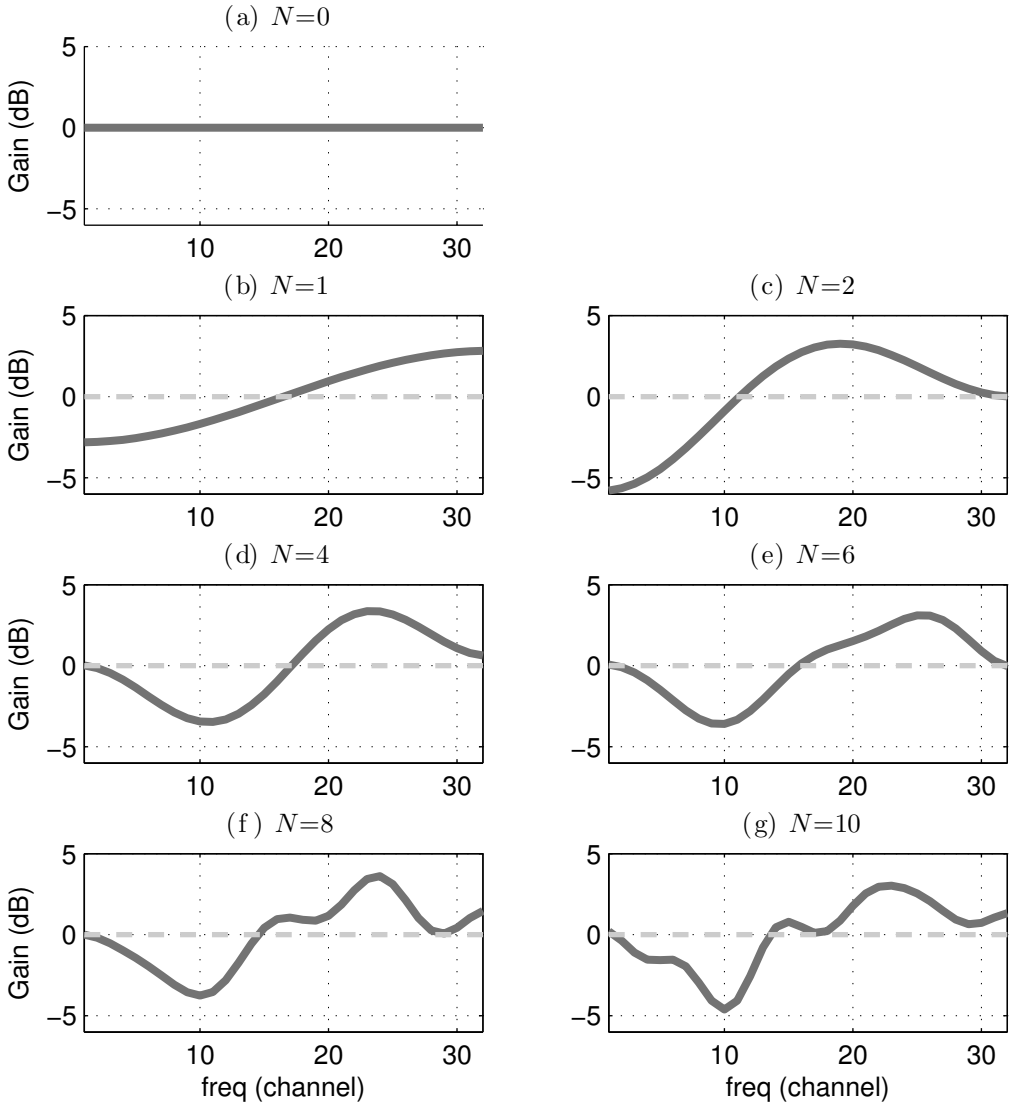

Figure 5.3: The optimal spectral shape in (dB) of GP-OPT modified speech for speech-shaped noise (SSN) using different number of cepstral coefficients, N, at a sentence level and averaged across a set of sentences covering the frequency range between between 50 and 8000 Hz. The $N = 0$ indicates the original unshaped case.

glimpses varies according to N. One might interpret this result as the consequences of introducing more spectral information in the spectral shape as N increases.

Table 5.3 and 5.4 demonstrates the performance of the GP-OPT system using objective intelligibility measures, *i.e.,* the STOI and the GP measures.

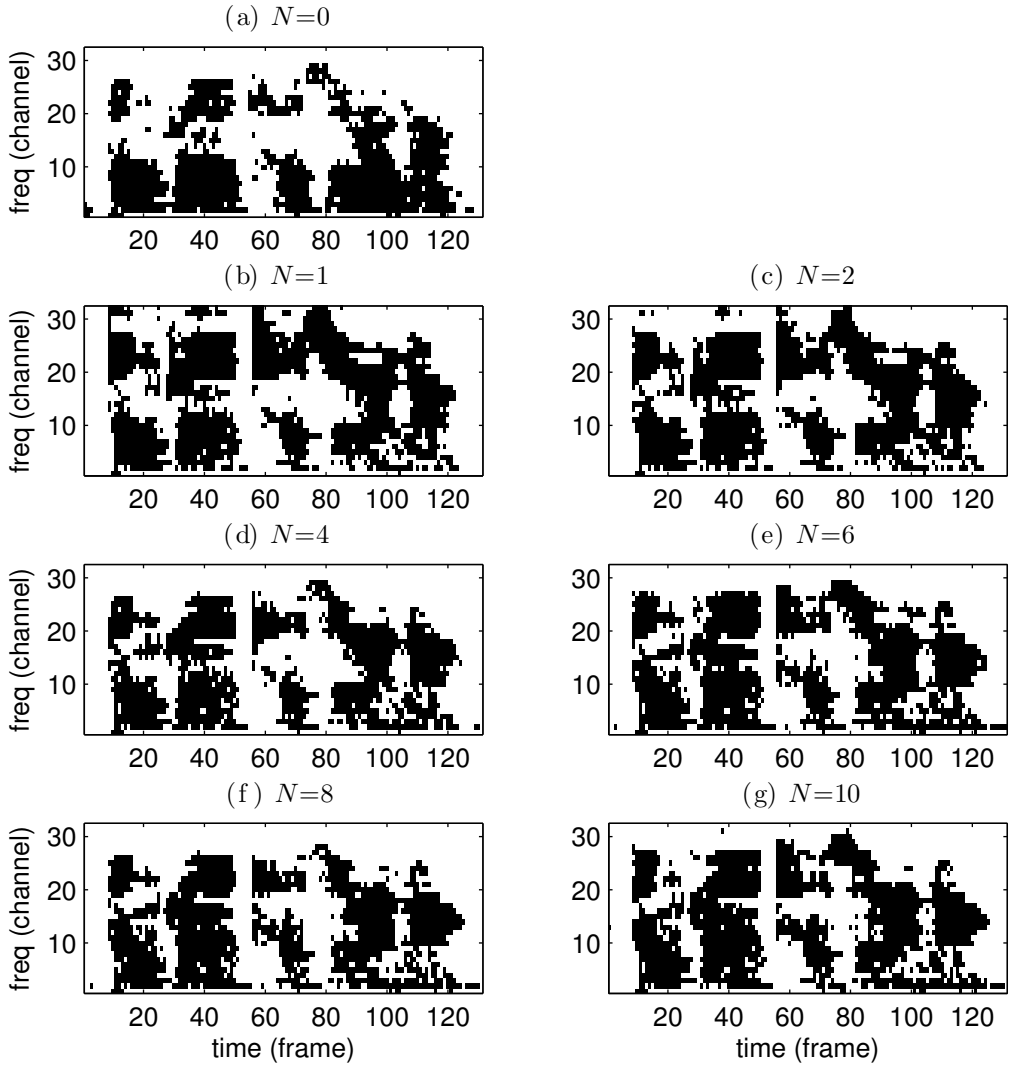

Figure 5.4: An illustration of example glimpse masks of GP-OPT modified speech using different number of cepstral coefficients N. The mask was generated for speech utterance of a male talker "bin green at k z now" and speech shaped noise (SSN) with equal overall rms levels and at a threshold, θ, of 0 dB, where white pixels indicate 0 and black pixels indicate 1, covering the frequency range between between 50 and 8000 Hz. The $N = 0$ indicates the original unshaped case.

Table 5.3 shows the results of the objective evaluation for the STOI measure at different setting of N, varying from 1 to 10 in SSN masker and at a range of SNRs. The results show an increase in predicted speech intelligibility relative to the unshaped case ($N = 0$). In general, when $N = 4, 6, 8$, the STOI scores show a better performance than

Table 5.3: The predicted intelligibility scores of GP-OPT modified speech using the Short-Term Objective Intelligibility (STOI) at a range of SNR for different number of of cepstral coefficients N in speech-shaped noise (SSN). The $N = 0$ indicates the original unshaped case.

	Number of Cepstral Coefficients						
SNR (dB)	$N = 0$	$N = 1$	$N = 2$	$N = 4$	$N = 6$	$N = 8$	$N = 10$
-9	0.51	0.52	0.52	0.53	0.54	0.54	0.53
-6	0.57	0.60	0.60	0.63	0.63	0.64	0.64
-3	0.63	0.68	0.68	0.70	0.71	0.72	0.72
0	0.72	0.75	0.76	0.78	0.78	0.79	0.79
3	0.79	0.82	0.82	0.83	0.84	0.85	0.85
6	0.86	0.87	0.87	0.89	0.89	0.89	0.89
9	0.91	0.92	0.91	0.93	0.93	0.93	0.93
Average Score	0.72	0.74	0.74	0.76	0.76	0.77	0.77

the $N = 0$ at mid noise level (*i.e.,* -6, and -3 dB SNR), while the results show almost the same performance at low noise level (*i.e.,* 0, 3, 6 and 9 dB SNR). At higher SNRs the differences between the scores across the cases tend to disappear.

Table 5.4 demonstrates the predicted intelligibility scores using the GP at a range of SNR for different number of of N. It is clear that there is an improvement of the average GP as N varies compared to the no-shaping case: ($N = 0$, $GP\% = 26.2$), ($N = 1$, $GP\% = 31.7$), ($N = 2$, $GP\% = 29.4$), ($N = 4$, $GP\% = 35.6$), ($N = 6$, $GP\% = 32.1$), ($N = 8$, $GP\% = 32.6$), ($N = 10$, $GP\% = 33.1$). The average GP score of the case $N = 4$ seems to outperform the remaining average GP at the different setting of N.

5.5.3 Performance analysis in stationary and non-stationary maskers

In the previous experiment, the performance was evaluated using the SSN masker at different numbers of cepstral coefficients varying from 1 to 10. The aim here is to study the impact of a different type of makers: stationary (*i.e.,* BMN) and non-stationary (*i.e.,* SSN) on the spectral shape averaged across SNRs and on the intelligibility improvement using the STOI.

Figure 5.5 shows an illustration of optimal spectral shape in (dB) of modified speech for SSN - *dark grey* - and BMN - *light grey* - at a sentence level and averaged across

Table 5.4: The predicted intelligibility scores of GP-OPT modified speech using the Glimpse Proportion (GP) with standard errors (σ_e) at a range of SNRs for different number of cepstral coefficients N in speech-shaped noise (SSN). The $N = 0$ indicates the original unshaped case. Here and elsewhere the error represent $\pm$ the standard error. The data represented as GP (σ_e).

	Number of Cepstral Coefficients						
SNR (dB)	$N = 0$	$N = 1$	$N = 2$	$N = 4$	$N = 6$	$N = 8$	$N = 10$
-9	11.4 (0.19)	12.5 (0.27)	11.9 (0.34)	15.6 (0.41)	12.7 (0.44)	12.8 (0.46)	12.0 (0.46)
-6	16.7 (0.17)	17.8 (0.24)	16.9 (0.31)	21.9 (0.39)	18.6 (0.45)	18.8 (0.49)	18.1 (0.50)
-3	20.5 (0.12)	24.0 (0.19)	22.8 (0.27)	28.9 (0.35)	24.2 (0.42)	25.8 (0.47)	25.1 (0.49)
0	26.6 (0.16)	31.0 (0.21)	28.4 (0.28)	35.1 (0.35)	31.2 (0.43)	32.0 (0.46)	33.5 (0.47)
3	31.1 (0.13)	38.3 (0.19)	36.7 (0.25)	43.8 (0.31)	39.9 (0.37)	40.8 (0.43)	40.1 (0.45)
6	37.7 (0.13)	46.8 (0.20)	43.1 (0.24)	50.2 (0.32)	47.7 (0.39)	47.2 (0.41)	48.9 (0.46)
9	43.9 (0.14)	53.0 (0.20)	50.3 (0.26)	58.7 (0.33)	54.4 (0.38)	55.0 (0.44)	56.5 (0.47)
Average Score	26.2 (0.15)	31.7 (0.22)	29.4 (0.28)	**35.6** (0.36)	32.1 (0.42)	32.6 (0.46)	33.1 (0.48)

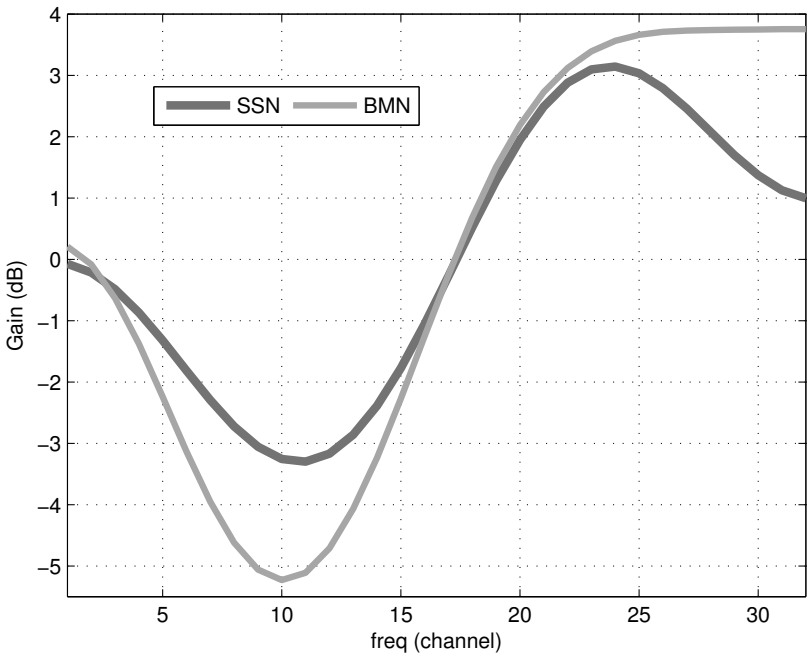

Figure 5.5: The spectral shape in (dB) of GP-OPT modified speech optimised for either speech-shaped noise (SSN) - *dark grey* - or babble-modulated noise (BMN) - *light grey* - at a sentence level and averaged across a set of sentences covering the frequency range between between 50 and 8000 Hz. Note that the number of cepstral coefficients, N, was set to 4.

the test set and across SNRs. Note that the number of cepstral coefficients, N, was set to 4, *i.e.,* (the optimal c_1, c_2, c_3, and c_4 were used to generate the average relative spectra). It is apparent that the spectral shape of BMN tend to take more energy from lower frequencies bands to the higher frequency bands. For example, the spectral shape of BMN differs by 2 dB at channel number 10 and 32 compared to SSN.

Additionally, The value of the GP is computed for each individual utterance and then averaged across the test set and noise type. The average intelligibility estimate of the ORG and GP-OPT method in SSN are: 26.2 % and 31.3 %, respectively. The average predicted intelligibility of BMN are: 31.6 % and 42.2 %, respectively. Even though the GP-OPT method was developed to optimise intelligibility in a stationary noise masker, *i.e.,* SSN, it also work well if it is optimised for a non-stationary noise masker, *i.e.,* BMN.

Figure 5.6 show the predicted intelligibility using the STOI at a range of SNRs. The left panel shows performance in the SSN masker, and the right panel shows performance in the BMN masker. The circles show performance of the GP-OPT modified speech as

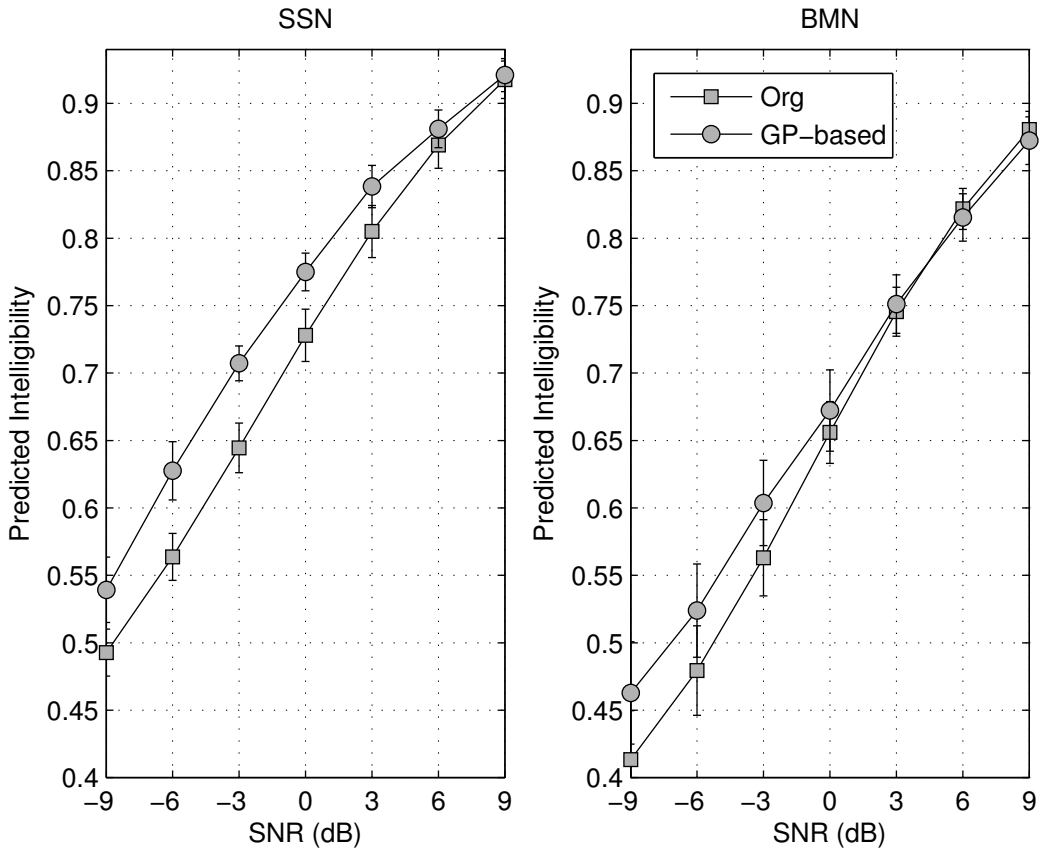

Figure 5.6: The predicted intelligibility scores of ORG and GP-OPT modified speech using the Short-Term Objective Intelligibility (STOI) with standard errors at a range of SNRs, where 0 represents maximum distortion and 1 represents minimum distortion to the speech signal. The left panel shows performance in the speech-shaped noise (SSN) conditions, and the right panel shows performance in the babble-modulated noise (BMN) conditions. Each point represents scores averaged across sentences for each SNR. The error bars represent ± the standard error.

a function of the SNRs. The square in the figure show performance of the ORG speech as a function of the SNRs. The error bars represent standard errors in each condition. Each point represents scores averaged across sentences for a certain SNR. It can be seen that the the trend of STOI scores of GP-OPT is improved in contrast to the ORG for both masker. In particular, the STOI predicts that the GP-OPT method improves the intelligibility at a high level of noise. For instance, The scores increase by roughly the same amount at −9, −6, −3, and 0 dB in the SSN condition. Furthermore, in the BMN condition the scores improve at −9, −6, and −3 dB. As the noise decrease (*i.e.,* 6, and 9 dB in both masker), STOI predicts that both the ORG and GP-OPT have the same

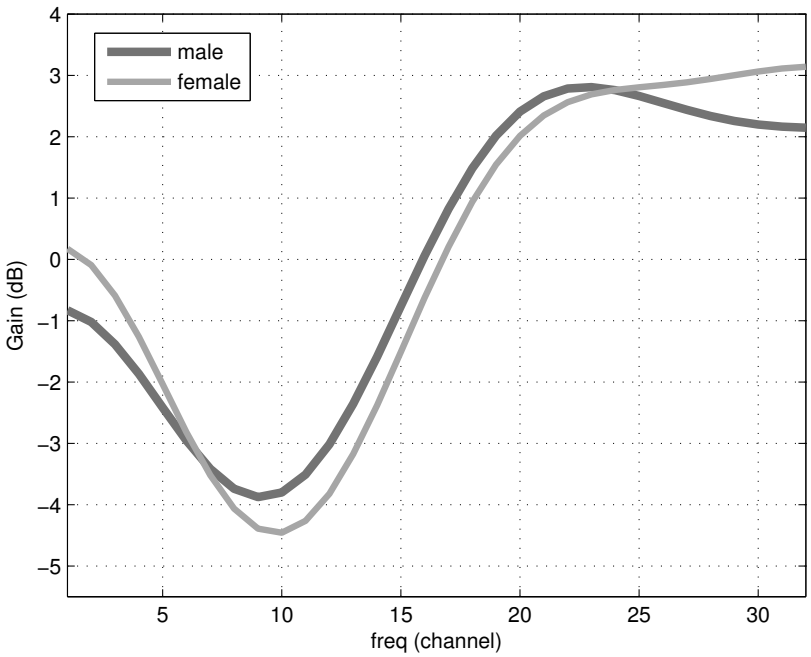

Figure 5.7: The spectral shape in (dB) of GP-OPT modified speech optimised for speech-shaped noise for males - *dark grey* - and females - *light grey* - at a sentence level and averaged across a set of sentences covering the frequency range between between 50 and 8000 Hz. Note that the number of cepstral coefficients, N, was set to 4.

level of intelligibility.

5.5.4 Performance analysis of gender differences

The aim of this analysis is to investigate whether males and females differs in their intelligibility benefit from the GP-based enhancement system. To achieve this aim, we study the average spectral spectral of males and females using a test set of 150 utterance each and were optimised for SSN. Further, the predicted scores using GP for GP-OPT verses ORG is studied for males and females.

Figure 5.7 shows the spectral shape in (dB) of GP-OPT modified speech averaged across a number of SNRs in SSN for male and female speakers separately. It is clear that the average spectral shapes are behaving similarly. Thus it can not be say that the males benefits from the enhancement system more than females or the vice versa.

Figure 5.8 shows the predicted intelligibility scores using the GP at a range of SNRs covering the frequency range between between 50 and 8000 Hz for male and female speakers separately for ORG in case (a) and GP-OPT in case (b), respectively. The

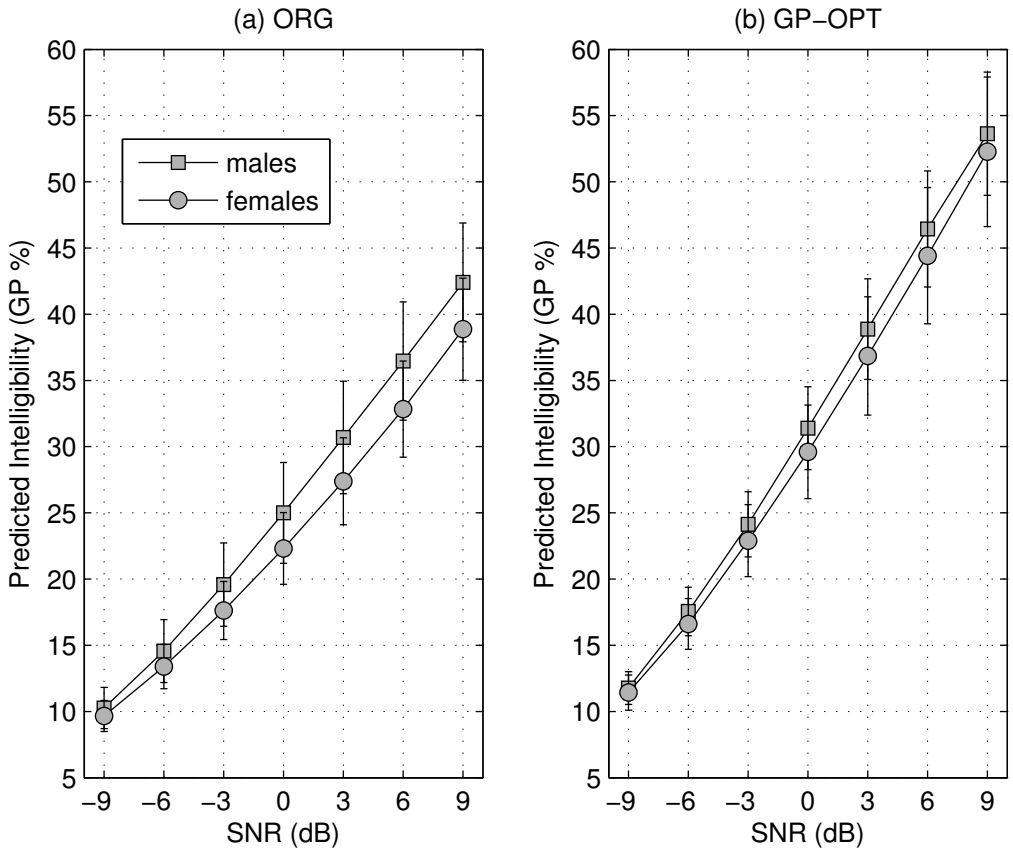

Figure 5.8: The predicted intelligibility scores of ORG and GP-OPT modified speech using the GP with standard errors optimised for speech-shaped noise at a range of SNRs for males and females. The $N = 0$ indicates the original unshaped case. Error bars represent $\pm$ the standard error.

average GP was computed for males and females individually and then averaged across the gender set for the SSN. The average GP for males and females in ORG case are: 25.57 % and 23.15 %, respectively. The average GP for males and females in GP-OPT case: 31.97 % and 30.58 %, respectively. It is clear that there is not differences between the estimated intelligibility.

This indicates that the GP measure used during the optimisation is less likely to consider the variability in gender specific data and hence the individual speakers' differences. This idea raises the possibility to improve the performance of the method by deriving a measure that accounts for the speakers variability in a speakers-dependent manner which will be examined in the following chapter.

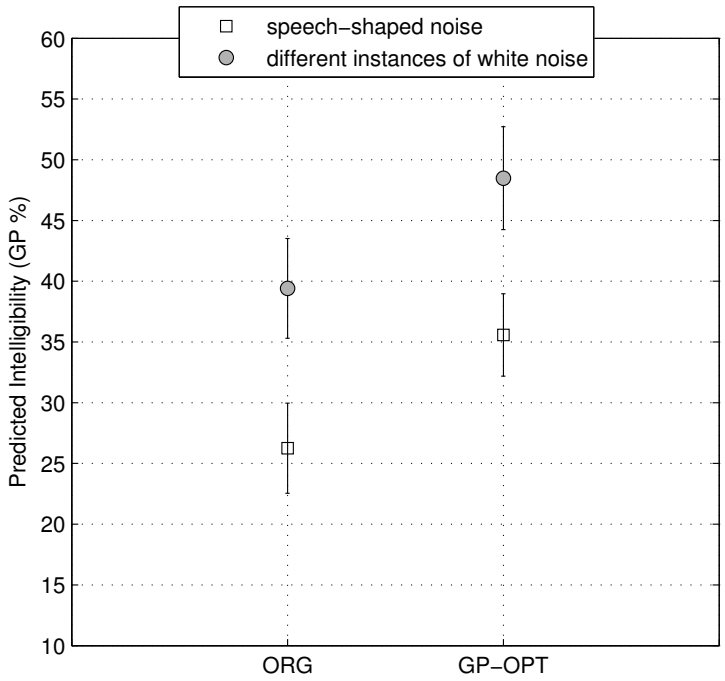

Figure 5.9: The predicted intelligibility using the GP measure with standard errors averaged across SNRs for ORG and GP-OPT (which optimised for speech-shaped noise) in the following noisy conditions: (i) *speech-shaped noise* denotes the same instance of speech-shaped noise used during the optimisation, and (ii) *different instances of white noise* denotes different instances of white noise NOT used during the optimisation.

5.5.5 Performance analysis using different instances of noise

The goal of this analysis is to study how the GP-OPT modified speech which has been optimised for stationary SSN noise, behaves in different instances of white noise, *i.e.*,it does not matched the one used during the optimisation. We use 100 instance of white noises. To evaluate the performance, the average intelligibility estimate is derived using the GP measure for both cases. In the different instances white noises case, the GP was derived for each mixture of the instance and then average across all the instances. and then across the test set.

Figure 5.9 shows the average predicted intelligibility using the GP measure with standard errors averaged across SNRs for GP-OPT and ORG in the following noisy conditions: (i) *speech-shaped noise* denotes the same instance of SSN used during the optimisation, and (ii) *different instances of white noise* denotes a different stationary noise type which was not used during the optimisation. It can be seen that the amount

of increase in GP of the ORG and GP-OPT speech was almost the same in both noise types (by 15%). It can be concluded that the GP-OPT works robustly even at using unmatched noise to the one used in the optimisation.

5.6 Human Listening Experiment

Objective speech intelligibility measures typically provides a performance indicator for a near-end intelligibility enhancement system. In spite of their beneficial role in designing the system, there is sometimes an inconsistency between the estimated results obtained from objective evaluation and the actual results obtained from subjective evaluation. For this reason, the intelligibility improvement of the near-end intelligibility enhancement systems should be validated by means of subjective listening tests.

In this section, the design and results of the listening test are reported. The primary purpose of the listening test was to quantify the intelligibility improvement of modified speech on native speech perception and to relate the findings to the predictive evaluation.

5.6.1 Participants

A total of 20 normal-hearing listeners participated in the experiment. All subjects were native English speakers. The listeners had no history of speech and/or language disorders, and their ages ranged from 20 to 30 years. All were paid for their participation. Ethics permission was obtained following the University of Sheffield Ethics Procedure.

5.6.2 Stimuli

The aim of the formal listening experiment was to evaluate the performance of four speech types; namely ORG, GP-OPT, SS and DIS-OPT (see Table 5.2)in SSN masker and over a wide range of SNRs: 3, 0, −3, −6 and −9 dB, and, therefore, to cover the range from low intelligibility to high speech intelligibility. The target stimulus was mixed with the masker during the testing procedure at a desired SNR level. The later entry, 'DIS-OPT', is not included in this chapter for presentation purpose. The results and development of the DIS-OPT method will be presented in the following chapter. The remaining included speech types are: ORG, GP-OPT, and SS.

The SS speech type applied the first subsystem of the Spectral Shaping and Dynamic Range Compression (SS-DRC) introduced in Zorila et al. (2012). The SS-DRC is a non-parametric pre-processing technique proposed to improve the intelligibility of speech in

presence of background noise. The technique preforms a spectral enhancement (spectral shaping) followed by temporal enhancement (dynamic range compression) which are combined in a cascaded way to form the SS-DRC system. It is not tuneable and it is noise-independent. It has been demonstrated that this approach is effective in a large-scale open evaluation of speech modification algorithms (Cooke et al., 2013a).

More specifically, each stimulus was generated using the SS system. It consists of two cascaded subsystems which are adaptive and fixed spectral shaping. The adaptive spectral Shaping is adapted to the probability of speech frame voicing and it applies formant enhancement. In the fixed spectral shaping, adaptive pre-emphasis filter is used in oder to control the distortion of speech modification by limiting the attenuation of high frequencies in speech signal. The outcomes of this technique is based on observations of formant enhancement in clear speech Hazan and Baker (2011) and spectral tilt reductions in Lombard speech Lu and Cooke (2008). The SS system is a frame-by-frame base analysis and synthesis system. For each frame, the envelope is computed using the Fast Fourier Transform (FFT), followed by applying the SS method. The modified speech signal is then re-synthesised using the concept of overlapping and adding procedure.

5.6.3 Procedure

The subjective intelligibility of the four speech types was tested in 5 noise conditions using a total of $13,600$ stimuli (4 enhancement conditions x 680 sentences (34 speakers x 20 utterances) x 5 conditions) divided into independent blocks of 136. The independent block was drawn at random, without replacement in which a single subject would hear 34 sentences from each entry into the 5 blocks (34 x 5 = 170 sentences in total). The subjects were assigned into blocks in which:

1. each subject heard one block of 136 (34 utterances x 4 enhancement conditions) sentences in each of the 5 noise conditions;

2. no subject heard the same sentence twice;

3. each noise condition was heard by the same number of subjects.

Subjects were instructed to identify the keywords (*i.e.,* **letter** and **digit**) spoken and type the heard keywords after listening to each stimulus that corresponded to one sentence. Once the subject had typed a response, the subsequent stimulus was presented automatically. Stimuli were presented once only, and subjects were not able to change the previous output. Null responses were not permitted. During a test, a subject was seated

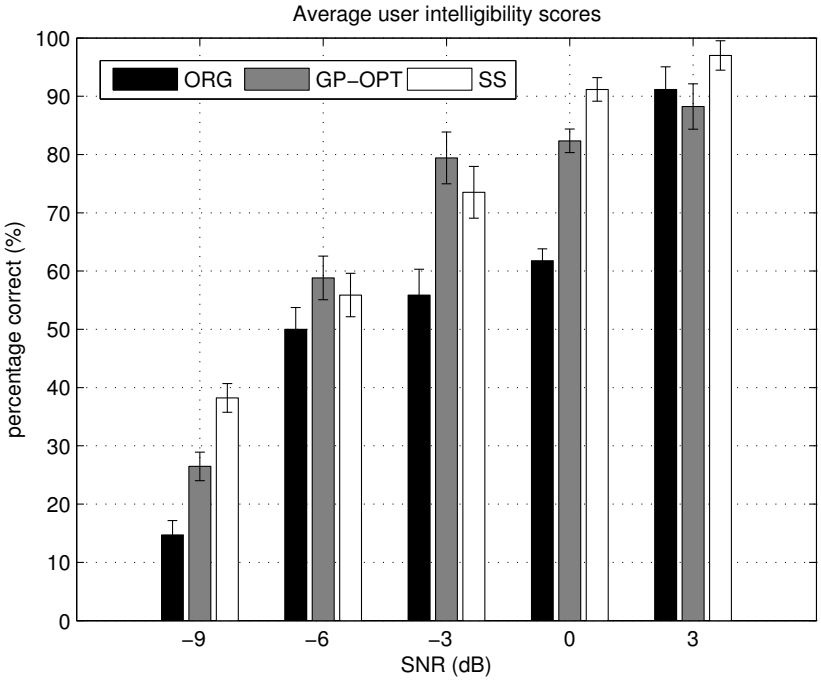

Figure 5.10: Percentage of correct identifications of both letter and digit in each speech type with the standard errors in the speech-shaped noise. The *black* shows performance in the ORG condition as a function of the signal-to-noise-ratio (SNR) used to produce the stimulus. The *gray* in the figure shows performance in the GP-OPT condition as a function of the SNR used to produce the stimulus. The *white* shows performance in the SS condition as a function of the SNR used to produce the stimulus. Note that all were presented at the same overall RMS level.

in an IAC single-walled acoustically-isolated booth. Test stimuli were generated using a web-based experiment system and then presented individually to a listener through headphones. A subject test took on average in 30-40 minutes. To familiarise them with the test procedure, subjects were given a training session at the beginning of the experiment by listening to and reporting on clean sentences. Subjects were unable to modify the output level with an average speech level of 68 dB SPL.

5.6.4 Results and discussion

Figure 5.10 shows the percentage of correct identifications of both letter and digit in each condition in the speech-shaped noise. The *black* shows performance in the ORG condition as a function of the signal-to-noise-ratio (SNR) used to produce the stimulus. The *gray* in the figure shows performance in the GP-OPT condition as a function of the

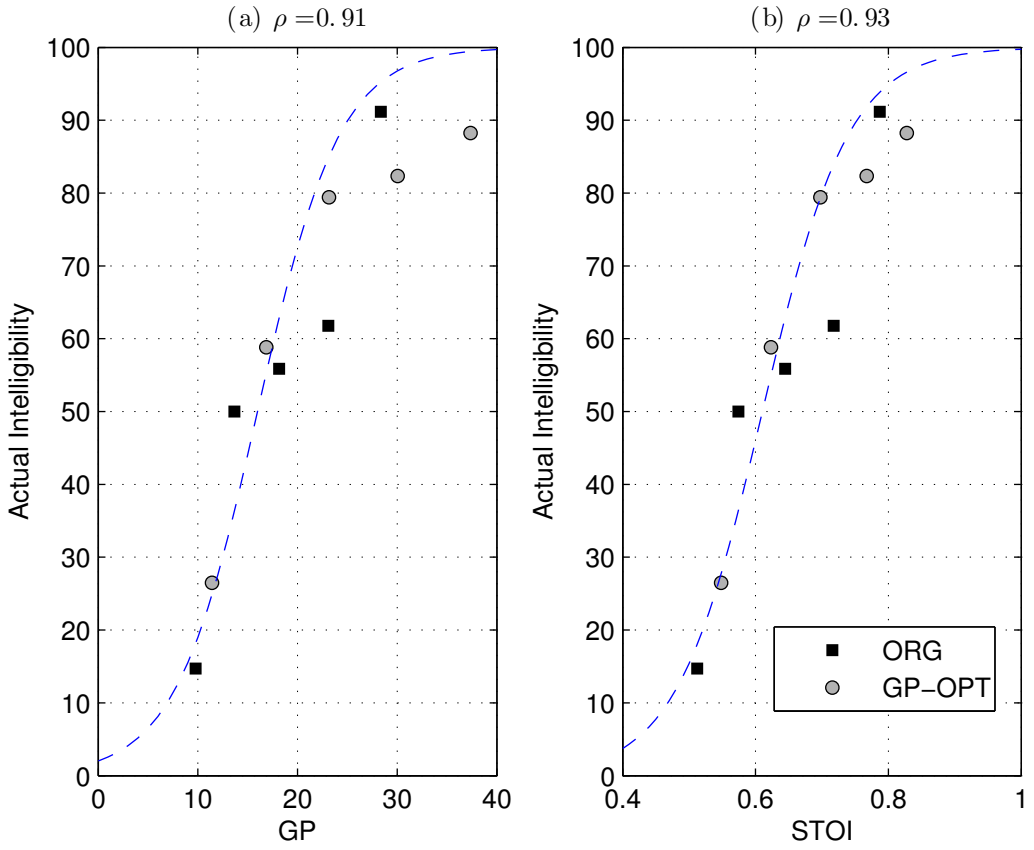

Figure 5.11: Scatter plots of STOI predictions versus actual intelligibility scores in (a) and GP predictions versus actual intelligibility scores in (b).

SNR used to produce the stimulus. The *white* shows performance in the SS condition as a function of the SNR used to produce the stimulus. The reported scores were computed as the average of percentage of correctly identified letters and digits. The error bars denote the standard error of the average score in each condition, calculated from the pooled data collected from all the subjects in the experiment.

Mean intelligibilities are 54.71 %, 67.01 %, and 71.17 % for ORG, GP-OPT, and SS across SNRs levels, respectively. The ANOVA results showed that the effects of modification type and noise level were significant ($F(2,8) = 7.75$, $p < 0.05$), and ($F(4,8) = 39.98$, $p < 0.05$), respectively. The Bonferroni as well as the Fisher LSD post hoc tests between the ORG and modifications conditions indicated that all means were significantly different ($p < 0.05$), whereas the differences between the GP-OPT and SS conditions were not significant ($p > 0.05$). The results show that GP-OPT and SS leads to better intelligibility compared to ORG mixtures in regardless of background noise.

The intelligibility for the GP-OPT in SSN masker is comparable to the SS.

Across modification types including the ORG, mean intelligibilities are 26.47 %, 54.90 %, 69.60 %, 78.43 % and 92.14 % for $-9, -6, -3, 0$, and 3 dB SNRs, respectively. The Bonferroni as well as the Fisher least significant difference (LSD) test indicated that the means were significantly different at -3, and 0 dB SNRs levels. The results show that the modification preforms well across the noise level, and it has a greater effect on the SSN background at at -3, and 0 dB.

Comparing the results from the modification conditions to those from the original conditions, it is apparent that performance for the GP-OPT and SS stimuli were very similar to one another, and that they were almost identical in shape to those obtained in the ORG condition at higher SNR level (i.e. 3 dB). This suggests that the application of GP-OPT and SS processing improved the intelligibility of a speech stimulus masked by SSN noise at lower SNRs level by roughly 25%-30% at -3 and 0 dB. Furthermore, this improvement appears to be highly relative to the results obtained by objective evaluation. Figure 5.11 illustrates the scatter plots of actual intelligibility scores (%) against (a) the STOI measure and (b) GP measure predictions for the GP-OPT speech and the unmodified ORG speech. Each point represents scores averaged across sentences and listeners for SSN noise. STOI and GP are used for prediction since they gave high correlation for the conditions GP-OPT and ORG ($\rho = 0.94$) and ($\rho = 0.91$), respectively, as shown in the left and right plots in Figure 5.11. The plot also shows the best fitting logistic function [2] to map the predictions scores to actual intelligibility scores.

5.7 General Discussion

The overall goal of this chapter was to design and implement a baseline for the pre-enhancement system. Specifically, we asked 'Can speech intelligibility actually benefits from an optimisation based closed-loop pre-enhancement ?' The results showed that the proposed approach does indeed improve speech intelligibility. The technique is based on increasing the number of unmasked S-T elements to reduce the amount of EM. However, the amount of that each S-T element contributes to the intelligibility is still uncertain because the used measure, GP, is an audibility measure that treats all S-T elements equally.

The findings suggest that using prior information about the speech and noise signals can improve the intelligibility for normal listeners, and that there are substantially

[2] computed using MATLAB glmfit function with normal distribution and logic link function.

stronger correlations between the objective and subjective evaluation of speech intelligibility. We posed at the start of this chapter a spectral modification technique defined using the first few cepstral coefficients. By choosing the first two coefficients the modification will act as a spectral tilt, and for our case, the first four coefficients act as a spectral shaper that is much more flexible in reallocating the energy between the frequency channels. That is, we picked a reasonable number of the cepstral coefficients as either too few numbers or too much is not an effective choice to make. It is clear as the number of cepstral coefficients increased the the details of spectral modification increased (as shown in Figure 5.3). Thus, the optimisation becomes more difficult since the searching space increased and the speech distortion increased too. This leads to a reduction in the intelligibility and therefore using a smaller number of the coefficients seems to be the reasonable step to take.

The spectral modification is usually based on the observation of Lombard speech where more energy is found at higher frequencies. It is typically seen as a generalisation of high-pass filtering, spectral tilt or centre of gravity changes in a closed-loop feedback system for the purpose of minimising the effect of EM in a known noise condition and depending on the spectral resolution (Cooke et al., 2014). This approach has been previously investigated by several researchers. Taal and Jensen (2013) implement a linear time-invariant filter that works by maximising an approximation of SII. Petkov et al. (2013) modify the speech spectrally by adjusting the filter gains of a simple filterbank to optimise an ASR based intelligibility measure. These two techniques are relatively low-resolution. Tang and Cooke (2012) propose to use a static spectral filter based on a much higher resolution using a 55-channel gammatone filterbank by maximising the GP measure. The technique proposed in this chapter is similar in concept to later technique in which the GP measure was maximised and a higher resolution using a 32-channel gammatone filterbank were implemented. However, It defines a spectral shaper using the cepstral coefficients to better tune the system to the optimal solution.

A second implication of our findings deals with the robustness of modified speech in a known masker. We found that speech becomes more robust to stationary masker even at lower SNRs. Knowledge of how much speech and noise available across S-T elements has played an important role in increasing the number of S-T elements of speech and decreasing those of noise at all tested SNRs. These S-T elements of speech, however, does not account for the crucial parts of speech responsible for overall speech intelligibility. Furthermore, it is worth noting that specific S-T elements that increase speech intelligibility for one speaker might not be effective with a different speaker. Thus

we need to establish a direct link between the S-T elements of speech and important parts of speech by accounting for talker-specific attributes. This idea will be investigated further in the coming chapter.

5.8 Chapter Summary

This chapter has presented a technique for optimising the spectral shaping applied by means of a closed-loop near-end intelligibility enhancement system using the GP measure.

It focuses on the situation where the speech signal and the noise background are known in advance, *e.g.,* imagine an audio engineer adding a speech track into the audio mix for a movie scene. The closed-loop approach is based on a parameterised spectral shaping that is applied to the speech signal and which effectively redistributes energy between frequency bands, *i.e.,* unmasking some speech components at the expense of others. The parameters of this enhancement are optimised with respect to a measure of energetic masking (*i.e.,* the GP). Results from human listening tests and objective tests showed that the proposed procedure improved the intelligibility of speech compared to the original at a range of SNRs in steady-stationary speech-shaped noise condition.

The GP, however, only considers the amount of masking in spectro-temporal representation and does not account for speakers' variabilities. The next chapter will develop an intelligibility measure that encodes information about which spectro-temporal speech elements are most informative in a speaker-dependent manner. This will be achieved by employing a statistical 'microscopic' intelligibility model.

Chapter 6

A Discriminative Microscopic Intelligibility Model for Spectral Modification

6.1 Introduction

The intelligibly metric employed in the previous chapter, the glimpse proportion 'GP' (Cooke, 2006), is an audibility measure that models energetic masking. It is computed as a proportion of unmasked noise-free S-T elements. These elements, however, may not carry an equal amount of information about important parts of speech. This issue can be illustrated by the visual analogy presented in Figure 6.1.

The top panel of the figure shows two separate images each of which contains a different letter: **P** and **B**, respectively, in a clean undistorted condition. Clearly, applying the GP metric in this case will result in 100% since no information has been lost. The middle and bottom panels has the same images but they are partially distorted. In both cases the distortion is an occlusion, *i.e.,* part of the image has been masked by a rectangle filled with dots. It is clear that applying the GP metric in both cases will result in 50% whether the distortion happens at the *upper half* of the letters in the images as seen in case (a) or at the *lower half* as seen in case (b).

Comparing the middle panel with the situation appearing in the bottom panel. It is clear that the images in the middle panel can be easily distinguished. Whereas, in the bottom panel, it is impossible to distinguish between the two images since the important region in both images that help us in discriminating between the two letters have been

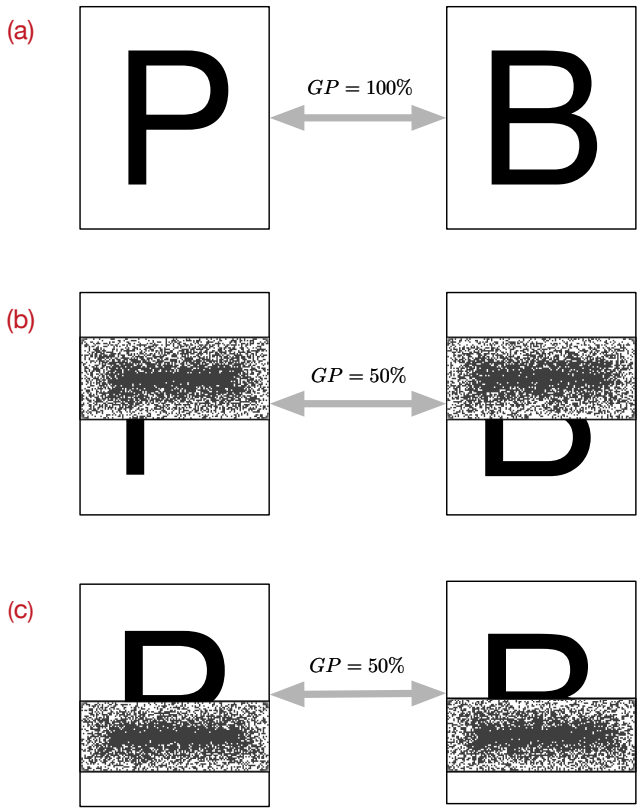

Figure 6.1: A visual analogy comparing how the GP model predict intelligibility of three cases each of which contains two images.

occluded. The GP measure does not account for the confusability of the information instead it accounts for the visibility of information after the occlusion. To address this problem, we need a discriminative based metric that account for the visible and inherent regions of the image. By applying this notion in audible information, we aim, in this chapter, to *define* an intelligibility metric that gives more weight to audible S-T elements that helps discriminate between confusable speech units, *e.g.,* confusable phonemes. We will then use this metric to *optimally enhance* the speech signal.

The measure described here extends our previous work in a number of respects. Firstly, our previous systems in chapter 5, used solely the missing data mask to derive the intelligibility metric. Here, we adopt a more logical approach in which the mask and an acoustic model of clean speech are used to deal with the missing data using marginalisation-based approach. Secondly, we define a way that turn the likelihood into a new discriminative intelligibility metric.

This chapter is organised into six sections. Section 6.2 is devoted to the computational development of the discriminative intelligibility metric. In Section 6.3, we show how to apply the metric in the closed-loop near-end intelligibility enhancement framework. Additionally, the implementation consideration is detailed in Section 6.4. The near-end intelligibly enhancement system is then objectively and subjectively evaluated in Section 6.5 and 6.6, respectively. The chapter is finally ends with a general discussion in Section 6.7 and a summary in Section 6.8.

6.2 Discriminative Microscopic Intelligibility Model Framework

In this section, we will introduce a framework for a new discriminative microscopic intelligibility model, referred to as 'DIS'. Its key idea is based on handling the missing data using the marginalisation approach which will be explained in Section 6.2.1. Then, the theoretical foundation of computing the DIS is described in Section 6.2.2.

6.2.1 Dealing with Missing Data

As stated previously, the first principle of dealing with the uncertain data is to estimate the missing data mask. In our application, the speech and noise are known a priori and therefore the mask estimation is a straightforward process (refer to Section 5.2 in the Chapter 5 for more details on how to derive the mask). This section explains how to deal with masked spectrogram statistically using missing data theory.

The sparseness and redundancy of S-T speech representations motivate the idea of dealing with missing data in the audible information and particularly in the context of speech recognition (Cooke et al., 2001). Handling missing or uncertain data is based on two main assumptions. The first assumes that there is a process that can estimate the reliable and missing elements. This process is referred to as 'missing data mask estimation'. The second is to match the reliable elements to models of clean speech. The missing or unreliable elements are thus treated by either marginalisation or imputation classification. Marginalisation-based approaches consider the range of possible values of the missing features. In the imputation-based approaches, on the other hand, the unreliable elements are replaced with estimates of the clean speech signal (Cooke et al., 2001). The marginalisation-based approach is advantageous, in our application, for two reasons. It has a more rational justification to be used here since it accounts for

all possible values for uncertainties, compared to the imputation-based approach that estimate a single value for uncertainty. Furthermore, in the latter approach, the missing features are replaced by clean speech estimates, after that recognition can proceed without modification of the recognition system, *e.g.,* (Raj et al., 2004). Therefore, in its straightforward form, missing data imputation is considered as feature enhancement stage prior to the recognition system, and hence it is hard to use in this application. In contrast, the marginalisation-based approach offers more flexibility to be used in different application, and therefore, it is applied as the fundamental of a discriminative intelligibility measure developed in this chapter.

The marginalisation-based approach to handle missing data is proposed in Cooke et al. (2001) to improve noise robustness in the context of speech recognition. Thus, the computational development is described as a classification problem in terms of the speech recognition. It is well known that hidden Markov models (HMMs) have been most widely-used in modelling the temporal behaviour of speech signals through a sequence of states, each of which is associated with a particular probability distribution of observations. Gaussian mixture models (GMMs) have been further used to estimate the probabilistic distribution of speech signals associated with each of these HMM states. For this reason, we will explain the fundamentals of the marginalisation-based approach in terms of HMM-GMM context. In this section, we briefly review the marginalisation-based approach to missing data as used for noise robust ASR.

The HMM-based speech recognition represents each speech unit by an HMM with a number of states. Each state, $\mathbf{q}$, is then modelled by a univariate Gaussian mixture distribution with diagonal covariance matrices over the units of $\mathbf{x}$ obtained from an input observation feature, X. Furthermore, two independence assumptions are usually made. First, the states $\mathbf{q}_i$ are conditionally independent of all other states taken into account the preceding state $\mathbf{q}_{i-1}$. Second, the observations are conditionally independent of all other observations taken into account the state $\mathbf{q}_t$ that yields $\mathbf{x}_t$.

Statistically speaking, given the sequence of acoustic observations $\mathbf{x}$ the aim now is to find the best state sequence $\hat{\mathbf{q}}$ as follows,

$$\hat{\mathbf{q}} = \underset{q}{\operatorname{argmax}}\, P(\mathbf{q}|\mathbf{x}) \tag{6.1}$$

By applying Bayes' rule, Equation 6.1 becomes:

$$\hat{\mathbf{q}} = \underset{q}{\operatorname{argmax}}\, \frac{P(\mathbf{x}|\mathbf{q})P(\mathbf{q})}{P(\mathbf{x})} \tag{6.2}$$

The probability of the acoustic observation sequence, *i.e.,* $P(\mathbf{x})$, stays constant within each state sequence, thus Equation 6.2 becomes:

$$\hat{\mathbf{q}} = \underset{q}{\operatorname{argmax}}\, P(\mathbf{x}|\mathbf{q})P(\mathbf{q}) \tag{6.3}$$

where the first part, $P(\mathbf{x}|\mathbf{q})$, denotes the observation likelihood and is named the acoustic model. The second part, $P(\mathbf{q})$ denotes the prior probability and is is termed the language model. For simplicity, the parameters in the likelihood and priors are assumed to be independent and therefore can be estimated separately.

Now, the obseravtion likelihood $P(\mathbf{x}|\mathbf{q})$ need to be determined. It is essential to note that in the missing-data strategy the HMMs are typically trained using clean speech and there is not a separate process for retraining in noise conditions. In real listening scenario, however, each observed acoustic feature vector, $\mathbf{x}$, is more likely to be distorted by background noise and thereby the calculation of the likelihood $P(\mathbf{x}|\mathbf{q})$ cannot be directly derived.

Now, Let us assume that there has been a prior segregation process that results in dividing the components of $\mathbf{x}$ into reliable features, $\mathbf{x}^r$, and unreliable features, $\mathbf{x}^u$. Therefore, each acoustic feature vector, $\mathbf{x}$, can be written as $\mathbf{x} = (\mathbf{x}^r, \mathbf{x}^u)$. In the $\mathbf{x}^u$, the speech information is treated as partially missing, and the challenge now is to preform classification with the remaining observed features, $\mathbf{x}^r$. In this chapter, we apply the marginalisation-based approach to classification. The approach computes the acoustic likelihoods by integrating over the range of all possible values of $\mathbf{x}^u$.

In the marginalisation-based (Cooke et al., 2001), the classification is based on the marginal distribution of the reliable features, $\mathbf{x}^r$, by integrating over the unreliable features, $\mathbf{x}^u$, in the state output distributions, as follow,

$$P(\mathbf{x}|\mathbf{q}) = \int P(\mathbf{x}|\mathbf{q})d\mathbf{x}^u = \int P(\mathbf{x}^r, \mathbf{x}^u|\mathbf{q})d\mathbf{x}^u \tag{6.4}$$

In GMM, the distribution $P(\mathbf{x}|\mathbf{q})$ is represented by a number of Gaussian distributions with diagonal covariance matrices. Exploiting the independency within each mixture component, the likelihood can be derives as follow;

$$P(\mathbf{x}|\mathbf{q}) = \sum_{k=1}^{M} P(k|\mathbf{q})P(\mathbf{x}|\mathbf{q}, k) \tag{6.5}$$

in which $P(k|\mathbf{q})$ denotes the weight for the mixture component k. Making the assump-

tion that the elements of $\mathbf{x}$ are independent, the acoustic likelihood in Equation 6.4 becomes:

$$P(\mathbf{x}|\mathbf{q}) = \sum_{k=1}^{M} P(k|\mathbf{q}) P(\mathbf{x}^r|\mathbf{q}, k) \int P(\mathbf{x}^u|\mathbf{q}, k) d\mathbf{x}^u \tag{6.6}$$

in which the distribution $P(\mathbf{x}^r|\mathbf{q}, k)$ denotes the univariate Gaussian distribution. The integral expression defines constraints on the range of values of the unreliable elements, $\mathbf{x}^u$.

The integral is carried out from $-\infty$ to $+\infty$ in a condition where unreliable features are entirely disregarded. Furthermore, assuming that the $\mathbf{x}$ represents a spectral energy vector and has a number of unreliable bands which are corrupted by noise. This reflects that the true range of speech energy in these bands have to be between zero and the observed energy $\mathbf{x}^u$. Thus, a constraint should be applied through bounding the range over which the unreliable components are integrated. The 'bounds constraint' is applied to Equation 6.6 to get the bounded marginal estimation of $p(\mathbf{x}|\mathbf{q})$ as follows,

$$P(\mathbf{x}|\mathbf{q}) = \sum_{k=1}^{M} P(k|\mathbf{q}) P(\mathbf{x}^r|\mathbf{q}, k) \frac{1}{\mathbf{x}^u} \int_{-\infty}^{\mathbf{x}^u} P(\mathbf{x}^u|\mathbf{q}, k) d\mathbf{x}^u \tag{6.7}$$

A consistent performance improvement has been reported when using the bounded marginals over using unbounded marginals (Cooke et al., 1997). When applying the multivariate Gaussian, the integral can be approximated by a vector difference of error functions in order to evaluate the bound marginals (Cooke et al., 2001) where the missing data mask is assumed to be discrete.

Moreover, the marginalisation-based methods have attracted many researchers relates to its ability of being more robust against data sparsity at low SNR's compared to the conventional imputation-based methods (Cooke et al., 2001).

6.2.2 Theoretical Foundation

In the preceding section, we have provided a brief description of how missing data method can be applied to noise-robust automatic speech recognition (ASR). The process typically involves solving two problems. The first is to identify S-T elements of reliable acoustic evidence referred to as a 'missing data mask', while the second requires modifying the ASR system in order to deal with the missing data. This section defines a discriminative microscopic intelligibility (DIS) measure, using the underlaying principles of the marginalisation-based missing data recognition. It uses an acoustic model, λ, to

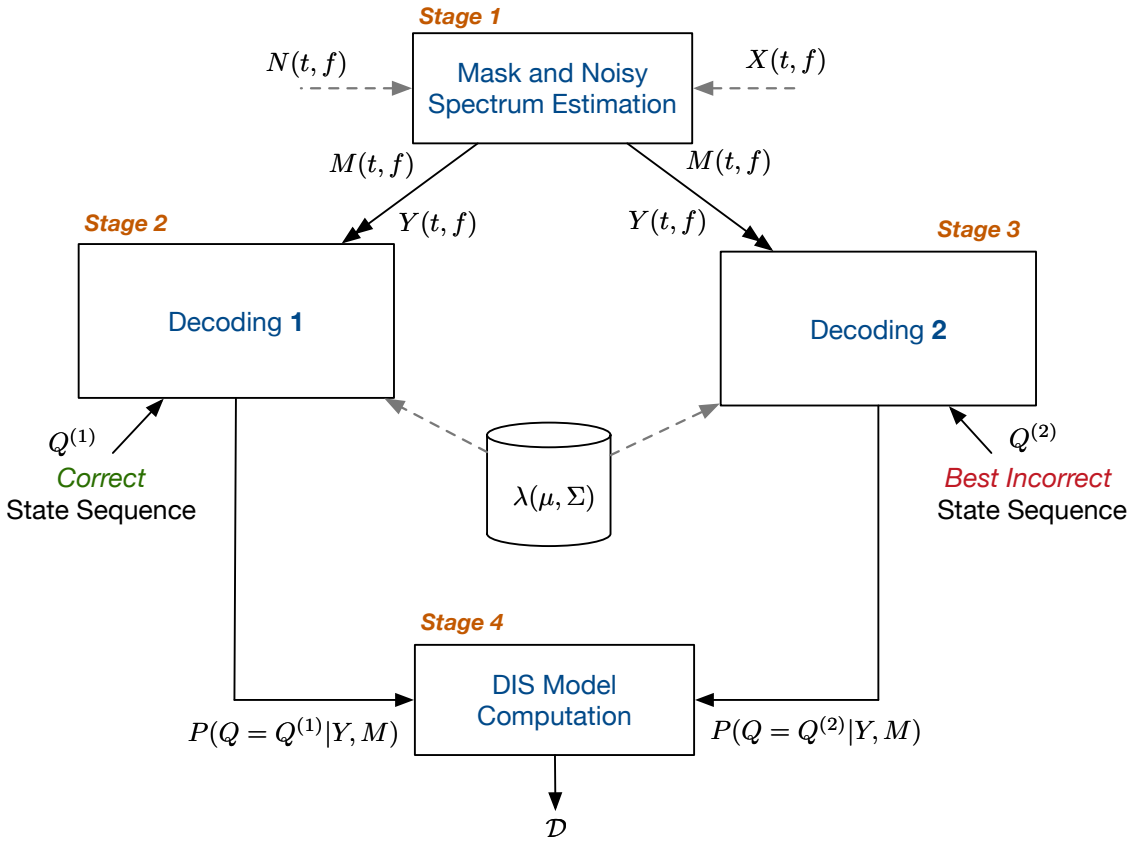

Figure 6.2: Block diagram of the overall structure of the discriminative microscopic intelligibility measure (DIS), where $X(t,f)$ and $N(t,f)$ are the S-T representation of speech and noise respectively, and $\lambda(\mu,\Sigma)$ is the acoustic speech model. The missing data mask, $M(t,f)$, and the noisy S-T representation, $Y(t,f)$, are derived using the $X(t,f)$ and $N(t,f)$. The $M(t,f)$ is defined as S-T elements wherein the speech power is greater than the noise power limited by some threshold value. Now the first decoding is performed using the *correct* state sequence, $Q^{(1)}$, and the likelihood is derived accordingly, while the second preformed using the *best incorrect* state sequence, $Q^{(2)}$, and the likelihood is computed accordingly. Then, both likelihoods are used to form the DIS, $\mathcal{D}$. Note, that the dashed grey lines denotes the main inputs and the boxes indicated the stages involved in the computation of DIS.

represent the speech by training HMMs on clean speech. Each HMM state is commonly represented as a GMM distribution with components having diagonal covariance. This form of HMM models are referred to as the GMM-HMM models. The development here follows the missing data work described by Barker et al. (2005). We introduce model and now we have got an opportunity to explore how much extra benefit we could gain from that specific information.

To start with, let Y denotes a sequence of noisy speech features $\{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ in which each $\mathbf{y}_t$ is a feature vector defining a spectral energy component at time t. The noisy spectrum can be approximated as the element-wise maximum of the speech spectrum, X and the noise spectrum, N, as follow

$$Y = \max(X, N),$$

Assuming further that the human listener is capable of making a perceptual segregation of the spectrum and estimating which frequency-bands are dominated by the background noise and which by the speech, *i.e.,* a missing data mask, $M(t, f)$, is known. In particular, the $M(t, f)$ represents a corresponding binary matrix representation that belonged to a sequence of frames $\{\mathbf{m}_1, \ldots, \mathbf{m}_T\}$, each frame being a vector of binary indicator variables denoting whether the corresponding S-T element is reliable or missing, '1' or '0' respectively.

Given the best underlying acoustic model state sequence $Q = \{\mathbf{q}_1, \ldots, \mathbf{q}_T\}$, the likelihood, $\mathcal{I}$, can be modelled as the probability that the phoneme is recognised accurately as follow:

$$\mathcal{I} = P(Q|Y, M), \tag{6.8}$$

By applying the missing data framework, the representation of the true signal, here X, is treated as a latent variable that we integrate over, as follows:

$$\mathcal{I} = \int_X P(Q, X|Y, M)\, dX, \tag{6.9}$$

By applying the Bayes' theorem, the probability in Equation 6.9 can be decomposed into:

$$\mathcal{I} = \int_X \frac{P(Q|X, Y, M)P(X|Y, M)}{P(Q)}\, dX, \tag{6.10}$$

To further simplify the problem, Equation 6.10 can be approximated as:

$$\mathcal{I} = \int_X \left(\frac{P(Q|X)P(X)}{P(Q)}\right).\left(\frac{P(X|Y, M)}{P(X)}\right) dX, \tag{6.11}$$

$$= \int_X P(X|Q)\frac{P(X|Y, M)}{P(X)}\, dX. \tag{6.12}$$

As mentioned earlier, the mask M defines a discrete segmentation of the spectrum X into 'reliable' bands X^r, *i.e.,* those where the speech is not masked and is thus directly observed, and 'unreliable' bands X^u where the speech is masked but known to have

energy less than the observed noisy value Y. The above equation can then be shown to be:

$$\mathcal{I} = \mathcal{C} \int_{X^u=-\infty}^{X^u=Y^u} P(X^r, X^u|Q)\, dX^u. \tag{6.13}$$

The constant $\mathcal{C}$ depends only on the observed spectrum, $Y(t, f)$, the prior $P(X)$ and the masking pattern $M(t, f)$. It can be ignored in missing data ASR because it is the same for all hypothesised states Q. However, in the case of near-end intelligibility enhancement application it cannot be ignored because the observed spectrum, Y is dependent on the parameters of the near-end intelligibility enhancement in which we optimise and hence the value of $\mathcal{C}$ will not be the same for all evaluations. Therefore, $\mathcal{C}$ is hard to evaluate accurately because it requires a good model of the clean speech spectrum prior, $P(X)$.

To avoid needing to evaluate the constant $\mathcal{C}$, the N-best missing data decoding is considered to get all alternative hypothesis phonemes using all possible grammar representations (transcriptions) of the input utterance. Thus, rather than computing the probability of the correct phoneme we can consider computing the ratio of probability between the correct phoneme and the most probable incorrect phoneme. This is well motivated because most listening errors will occur not due to accumulated probability of all the incorrect classes but just due to confusing the correct class with the class that is most acoustically similar, *e.g.,* /b/ versus /v/ or /m/ versus /n/. The DIS score is defined as a ratio where the numerator HMM is the correct states sequence, $Q = Q^{(1)}$, and the denominator HMM is a recognition model containing a chosen best incorrect sequence, $Q = Q^{(2)}$, from all possible hypotheses sequences, as follows:

$$\mathcal{D} = \frac{P(Q = Q^{(1)}|Y, M)}{P(Q = Q^{(2)}|Y, M)} \tag{6.14}$$

$$= \frac{\mathcal{C} \int_{-\infty}^{Y^u} P(X^r, X^u|Q = Q^{(1)})\, dX^u}{\mathcal{C} \int_{-\infty}^{Y^u} P(X^r, X^u|Q = Q^{(2)})\, dX^u}. \tag{6.15}$$

where the constant $\mathcal{C}$ can be cancelled from the numerator and denominator. It should be noted that the DIS model is differed from the measure proposed by Petkov et al. (2013) where the probability of correct recognition of the transmitted message is optimised.

The stages involved in computing the DIS measure can be better illustrated by four main stages as shown in Figure 6.2. The first stage deals with estimating the missing data mask, $M(t, f)$, and the noisy spectrum, $Y(t, f)$. The second and third stages applies missing data decoding using $Q = Q^{(1)}$ and $Q = Q^{(2)}$, respectively. This is required since the probabilities in the N-best do not always match those in the 1-best decoding. This is simply because the implemented technique is a fast 'approximation' to the true

N-best and thereby it can underestimate the true probability of the N-best hypotheses (Schwartz and Austin, 1990). Although errors seem to be small, they may confuse the optimisation algorithm. Thus the easiest solution is to use the N-best list only in order to learn what state sequence produces the best scoring incorrect sequence, and then to score it accurately by forcing the decoder through this sequence. Finally, the outputs of the decoding (1 and 2) in second and third stages are used as inputs to the forth stage to derived the DIS measure proposed in this chapter.

6.3 Optimising Intelligibility

The development of the DIS model was detailed in the previous section. This section brings the the DIS model and the spectral modification (described earlier in Section 5.3 from Chapter 5) together in the optimisation based near-end intelligibility enhancement framework. Details are given below.

Precisely, using missing data approach, for a given noise, one could compute the probability for a target utterance given the pre-enhanced signal. The enhancement is then performed by finding the spectral shaping that maximises this probability, *i.e.,* maximising the probability that the model 'hears' the word correctly. Thus, the system can now be optimised by searching for the optimal cepstral parameters $\hat{\mathbf{c}}$, where

$$\hat{\mathbf{c}} = [\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_n]^T,$$

that maximise the $\mathcal{D}$ measure for a given speech x, noise n, and speech model λ, *i.e.,* the optimal parameter values are given by,

$$\hat{\mathbf{c}} = \underset{\mathbf{c}}{\operatorname{argmax}}\, \mathcal{D}(x, n, \lambda; \mathbf{c}), \tag{6.16}$$

$$= \underset{\mathbf{c}}{\operatorname{argmax}} \left(\frac{P(Q = Q^{(1)}|Y_c, M_c)}{P(Q = Q^{(2)}|Y_c, M_c)} \right), \tag{6.17}$$

$$= \underset{\mathbf{c}}{\operatorname{argmax}} \left(\frac{\int_{-\infty}^{Y_c^u} P(X_c^r, X_c^u|Q = Q^{(1)})\, dX_c^u}{\int_{-\infty}^{Y_c^u} P(X_c^r, X_c^u|Q = Q^{(2)})\, dX_c^u} \right). \tag{6.18}$$

Optimisation is performed using the Nelder-Mead Direct Search method (Lagarias et al., 1998), described earlier in Section 4.4.3. Figure 6.3 illustrates how the algorithm works. For each iteration the following steps are performed: (i) the speech is spectrally shaped according to the current parameter values, $\mathbf{c}$; (ii) the noisy representation and mask

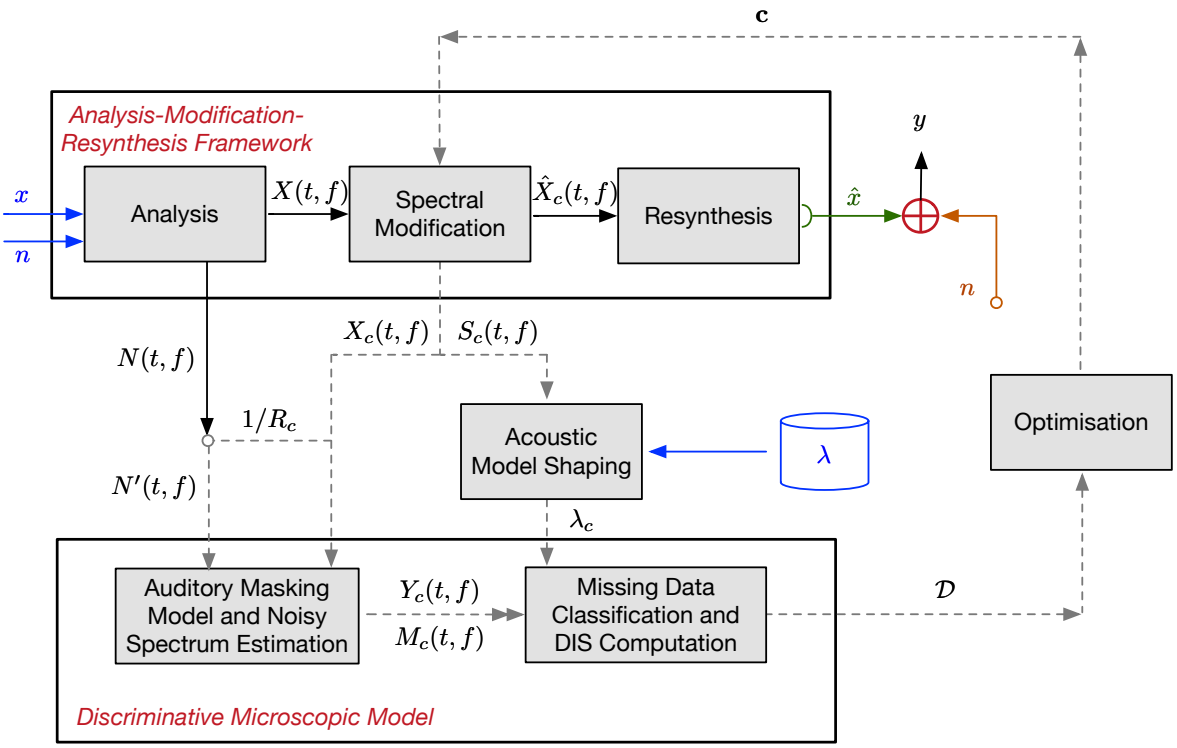

Figure 6.3: A schematic diagram of optimisation process for the DIS-based spectral modification system, where x and n denote the input speech and noise signals. The $X(t, f)$ and $N(t, f)$ are the spectrums, *i.e.*, the Spectro-Temporal Excitation Pattern (STEP), of x and n, respectively. Further, λ and λ_c represents the speech model before and after shaping. The R_c denotes the root-mean-square (RMS) after applying the optimal spectral shaping to speech spectrum. The $X_c(t, f)$ and $N'(t, f)$ denote the spectrums of the shaped speech and the re-normalised noise by factor $(1/R_c)$, both spectrums are used to compute the missing data mask $M_c(t, f)$ and the noisy spectrum $Y_c(t, f)$. The $\mathbf{c}$ represents the optimal cepstral coefficients after the optimisation and $\mathcal{D}$ is the DIS model. Solid arrows indicate fixed input to the optimisation process, whereas the grey dashed arrows indicate iterative process of optimisation.

are computed; (iii) the spectral shaping is applied to the HMMs; (iv) the missing data speech recogniser is first run with a grammar fixed to the correct utterance, and then again with the full grammar for the speech corpus being used and with N-best decoding ($N = 2$); (v) the difference between the log probabilities of the correct and best incorrect decoding is returned as the function value to be optimised.

6.4 Implementation Considerations

The necessary probabilities can be estimated using HMMs to model the sequence data without changing any of the fundamentals of the theory. The state sequence: Q, becomes the most probable state sequence through the known correct *word* sequence; the competing state, $Q^{(2)}$, becomes the most probable sequence through the most probable incorrect *word* sequence. They can both be estimated using N-best decoding with N set to 2, as explained in the previous section.

The binary mask $M_c(t, f)$ is computed from the known speech and noise signal using the same criterion used to define the S-T glimpses in the previous chapter (see Section 5.2 for more details), as follow;

$$M_c(t, f) = \mathcal{H}\left(X_c(t, f) > (N'(t, f) + \theta)\right). \tag{6.19}$$

in which the $X_c(t, f)$ and $N'(t, f)$ denote the STEP representation of the pre-enhanced speech and scaled noise (by a $1/R_c$ factor and R_c denotes the root-mean-square (RMS) after applying the optimal spectral shaping to speech spectrum) at time frame t and frequency band f, respectively. $\mathcal{H}(.)$ is the Heaviside step function counting the number of 'glimpses' which meet the local audibility criterion θ. Mask entries equal to 1 indicate the observation is present (*i.e.,* unmasked) and 0 indicates missing (*i.e.,* masked).

Additionally, it is assumed that spectral shaping does not alter the intelligibility of the clean signal, *i.e.,* it is assumed that listeners can adapt their internal models of speech to fit the shaped signal, known as *perceptual constancy* (Watkins, 2005). This is a common experience because the human listeners have the ability to cope with variability in natural environments, and thus the the auditory system achieves perceptual constancy even in complex listening conditions.

To model this the HMMs are spectrally shaped to match the shaping being applied to the speech. Given that the shaping, $S_c(f)$, is a fixed offset to the log spectrum, the models, λ, can be adapted by simply adding the same offset to the mean vectors of every Gaussian mixture component of every HMM model state. The process is known as 'acoustic model shaping' as shown in Figure 6.3 where

$$\lambda_c\left(\mu_c, \Sigma\right) = \lambda_c\left(\mu + S_c(f), \Sigma\right),$$

6.5 Objective Evaluation

In the previous section, we propose a novel technique for optimising the spectral shaping parameters using the DIS intelligibility measure in the near-end intelligibility enhancement system. The approach is designed to be used in situations where both the speech and noise signals as well as a statistical model of the speech are known a *priori*.

In this section, we carry out two analyses studies in order to evaluate the effectiveness of the discriminative-based enhancement system, 'DIS-OPT'. The first analysis, in Section 6.5.2, investigates the possible benefits that can be achieved from applying not-knowing and knowing the talker scenarios. This analysis is then extended in Section 6.5.3 to study the impact of knowing the talker scenario on each individual talker of Grid corpus. Finally, the overall performance of the system is assessed in Section 6.5.4 using different objective intelligibility measures (OIMs). Formal listening tests are also planned to verify the predictions made by the objective intelligibility models, that will be presented in the following section.

6.5.1 Experimental setup

In this section, we present the common experimental setup implemented in this chapter.

Speech and noise materials

Experiments were performed using speech data taken from the Grid corpus (reader refers to (Cooke et al., 2006) for more information about the Grid corpus). The corpus consists of 34 native English talkers (18 male and 16 female) speaking simple 6-word command sentences from a fixed grammar. There are 1000 utterances recorded from each speaker sampled at 25 kHz. A test set contains 600 utterances (around 20 utterances per talker) were randomly drawn from the Grid corpus for analysis purposes.

A stationary speech shaped noise (SSN) was used for evaluation in this chapter. It was generated, as detailed in Chapter 5, by filtering white Gaussian noise through a 100-order all-pole filter, the long-term average spectrum of this noise was approximated to match that of the Grid speech material. To produce noisy mixtures, the SSN masker were artificially added to the 600 test utterances at a range of global SNRs: -9 to 9 dB with an interval of 3 dB. Utterances were normalised to have equal RMS energy and sampled at 25 kHz.

Table 6.1: Speech types used for the evaluation in this chapter.

Type	*System*	*Required Knowledge*
ORG	Original unmodified speech	-
GP-OPT	GP-optimised speech using spectral shaping modification technique (technique developed in Chapter 5)	Speech and noise signals
DIS-OPT	DIS-optimised speech using spectral shaping modification (DIS measure developed in the current Chapter)	Speech and noise signals, a model of clean speech
SS	modified speech using spectral shaping developed by Zorila et al. (2012) (*i.e.,* only the first part of their system is implemented here)	Speech signal

Speech types

Four speech types are considered for evaluation in this chapter which are three types of modified speech and the original (unmodified) speech 'ORG', as shown in Table 6.1. The three of modified speech types are: (i) the GP-optimised speech 'GP-OPT' that described in the previous chapter (see Section 5.4), (ii) the discriminative-optimised speech, referred to as 'DIS-OPT', and finally (iii) as a reference system, modified speech using spectral shaping developed by Zorila et al. (2012), 'SS', which is used in the listening test which will be described in the following section. We used the test set to produce the modified speech types.

The DIS-OPT modified speech is generated in a the same procedure used to get the GP-OPT modified speech. However, the optimal cepstral parameters are different. To produce the DIS-OPT modified speech, the optimal cepstral parameters are first estimated using the optimisation described in Section 6.3. Then the spectrum of speech shaped in the analysis-modification-resynthesis framework. In particular, the speech signal is filtered using a bank of 32 gammatone filters with centre frequencies spread evenly on the ERB scale between 50 and 8000 Hz with filter bandwidths matched to the ERB of human auditory filters. Then, the envelope of each gammatone filter output is computed. This envelope is then smoothed by a first-order low-pass filter with an 8 ms time constant. Then, the smoothed envelope is down-sampled to 100 Hz. Following downsampling, the amplitude envelope is logged to turn the amplitude into

the log-energy domain. The spectrum is then shaped by applying a band-dependent scaling using the optimal cepstral parameters to the gammatone filter outputs before re-summing them to form the pre-enhanced signal, $\hat{x}$. Re-synthesising is then applied to generate the spectrally shaped speech signal. Care needs to be taken when summing the bands to compensate for band-dependent phase delays introduced by the analysis (for more details see Section 4.2.2). After re-synthesising the spectrally shaped signal is scaled such that the global signal energy remains unchanged before and after spectral manipulation. The result $\hat{x}$ signal was mixed with the masker at a desired SNR level.

For each noisy utterance the optimal speech spectral shaping parameters, $\hat{\mathbf{c}}$ are determined using the Nelder-Mead Direct, *i.e.,* 'simplex' Search method. This simple optimisation approach only requires evaluation of the function and does not require derivatives. It was found to converge quickly when the shaping vector $\mathbf{c}$ was initialised to $\mathbf{0}$ so there was no need to use more sophisticated algorithms. Further, the number of cepstral coefficient was $N = 4$.

Acoustic speech models

To train acoustic speech models[1], a 17,000 utterance training set was provided containing 500 utterances of each of the 34 Grid talkers. Each of the 51 words in the Grid vocabulary is modelled with a left-to-right HMM with 2 states per phoneme, *i.e.,* models vary in length from 2 states to 10 states. Precisely, the spoken letters, 'at', 'by' and 'in' used 4-state models while all remaining words employed 6-state models apart from 'again' and 'zero' (8 states) and 'seven' (10 states). The emission probability for each state is represented by a 7 component diagonal-covariance GMM. The language model is fixed based on the Grid syntax, *i.e.,* of the form '<command> <colour> <preposition> <letter> <number> <adverb>'. Training is carried out in two stages. We first train a speaker-independent (SI) model from a flat start using the full 17,000 utterances training set. We further derived a speaker-dependent (SD) model for each of the 34 speakers by running further parameter re-estimations using just the target speaker training data.

[1]The HMM-speech models were constructed as in the PASCAL CHiME speech separation and recognition challenge (Barker et al., 2013; Cooke et al., 2010).

6.5.2 Statistical analysis of known and not-known talker scenarios using average relative spectra

The fundamental task in the design of the near-end intelligibility enhancement system here is to find the optimal spectral shaping parameters that are thought to unmask elements of the S-T representation and provide good discriminability. The choice is based mainly on optimisation procedure using knowledge about speech and noise signals, and an HMM-based model of speech.

The question to be answer here in this analysis is whether knowing the speaker allow the system to produce better results using average relative spectra. The 'relative spectra' is computed as the log differences between the average spectral envelopes calculated using all frames of the ORG and modified speech. It is obtained of ORG speech over modified speech. The resulting relative spectra are then averaged and represented in (dB). We illustrate a broad analysis of a set of utterances that contains roughly 20 utterances per talker (20×34). To test this, we assume that the target utterances were spoken by one talker from the talkers encountered in the training set, but two different configurations were employed: (i) *(not-known talker scenario):* the HMMs-based SI model were implemented, and (ii) *(known talker scenario):* the HMMs-based SD model corresponding to the talker who spoken the target utterance were used. To further investigate the amount of changes introduced in the average relative spectra, we define the degree of change (DC) factor. The DC factor is computed as the average absolute change of the average relative spectra in dB.

Figure 6.4 shows the average relative spectra for DIS-OPT averaged across all frames and all talkers in the Grid corpus in the SSN masker for both *known* and *not-known* talker scenarios compared to ORG. Furthermore, Table 6.2 demonstrates The degree of change and the standard deviation of error for average relative spectra for DIS-OPT of all talkers in the Grid corpus in the SSN masker for both *known* and *not-known* talker scenarios. It can be seen that, for the known case, the amount of DC gain is larger than that of the not-known case. Thus, it could be inferred that the system behaviour benefits from the extra knowledge used about the talker during the processing.

Figure 6.5 and 6.6 illustrate the average relative spectra of a *male* and *female* talkers, respectively, (using roughly 20 utterance each) (left panels). Example masks (right panels) were also shown of a target utterance spoken by the male talker 'bin green at k zero now' and of a target utterance spoken by the female talker 'bin green by a zero now' for known and not-known talker scenarios for DIS-OPT compared to ORG in SSN

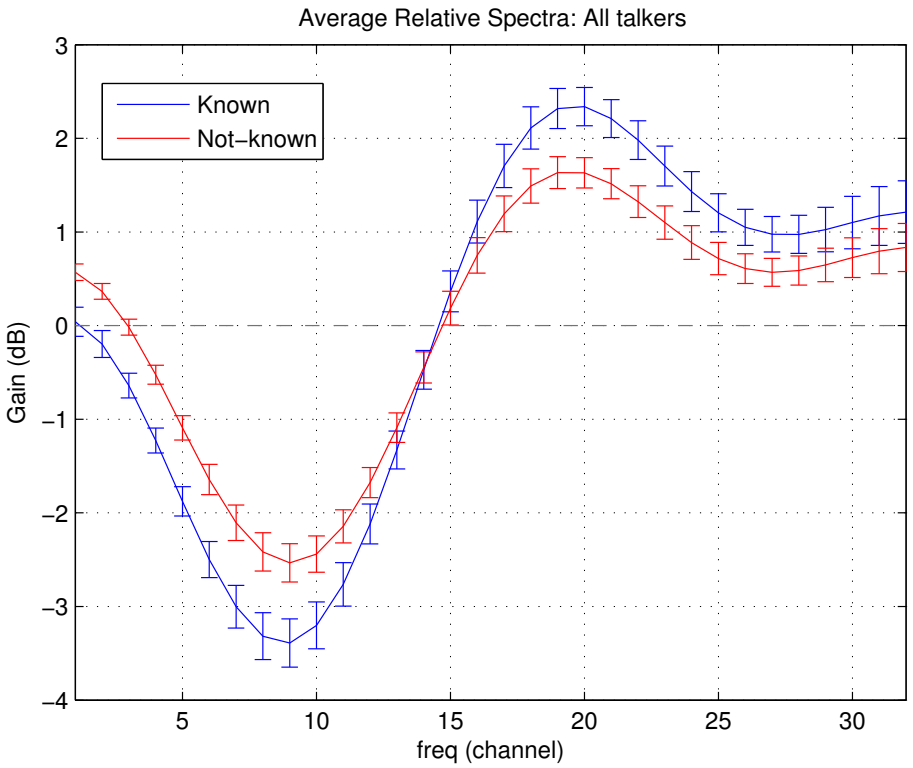

Figure 6.4: The average relative spectra for DIS-OPT averaged across all frames and all talkers in the Grid corpus in the SSN masker for both *known* and *not-known* talker scenarios, covering the frequency range between between 50 and 8000 Hz. The dashed line represents the original unmodified case (ORG).

Table 6.2: The degree of change (in dB) and the standard deviation of error for average relative spectra for DIS-OPT of all talkers in the Grid corpus in the SSN masker for both *known* and *not-known* talker scenarios.

Talker Scenario	*Degree of Change (DC)*	*Standard Error* (σ_e)
Not-known	1.13	0.17
Known	1.62	0.21

at 0 dB SNR. In the male speaker case, it can be seen that in the known configuration scenario the energy is moving from the middle to the higher frequency channels, whereas in the not-known configuration scenario a small amount of energy is reallocated. Moving to the female talker case, it is clearly evident that the S-T energy regions are taken from the lower to the middle and higher frequency channels. However, in the not-known configuration scenario, it is apparent that the amount of energy moved from the lower

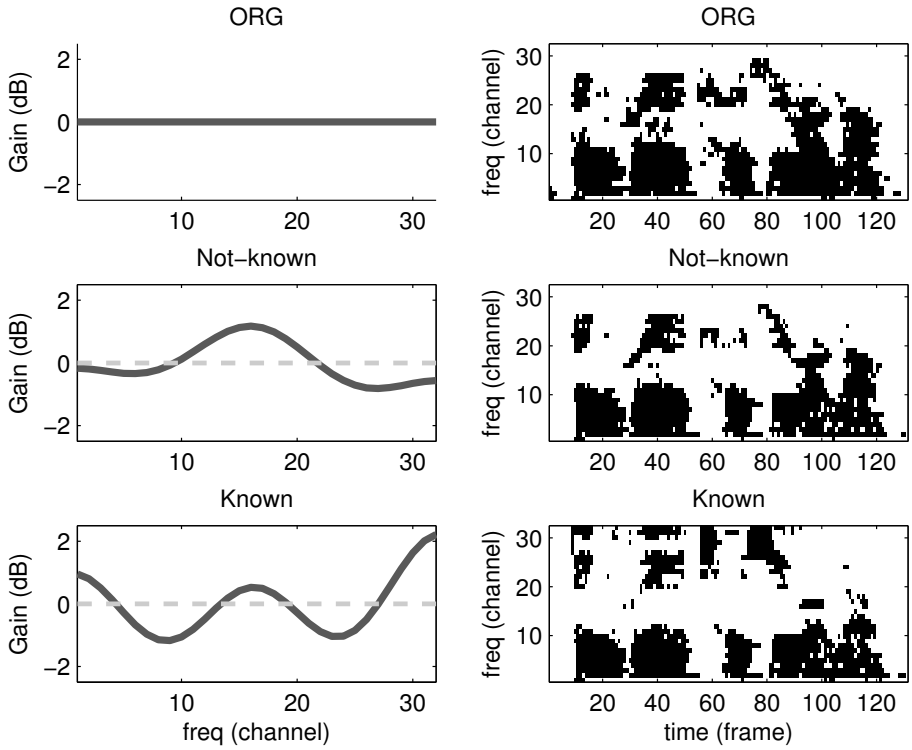

Figure 6.5: The average relative spectra of a *male* talker of roughly 20 utterance (left panels) and example masks (right panels) of a target utterance 'bin green at k zero now' for known and not-known talker scenarios for DIS-OPT compared to ORG in SSN at 0 dB SNR: (i) *Not-known* denotes using a speaker-independent (SI) acoustic model in the optimisation process; and (ii) *Known* denotes using a speaker-dependent (SD) acoustic model in the optimisation process, covering the frequency range between between 50 and 8000 Hz. Note, the male talker is (id = 10).

frequency channels is high in contrast to the known talker scenario.

In general, using the SD acoustic model seem to behave in a similar pattern for both male and female example talkers which introduced a more balanced spectral shape. Moreover, in the average relative spectra ensure more energy is moved to the higher frequency channels. This can be explained in terms of the empirically-observed behaviour of speech produced in noise,*i.e.,* the Lombard speech (Picheny et al., 1985; Van Summers et al., 1988) such as reducing spectral tilt to boost the high frequencies (Junqua, 1993; Lu and Cooke, 2008; Van Summers et al., 1988). The average relative spectra thus maintains similar objective in the SSN maker. Thus, the following experiment investigates the average relative spectra for each individual talker across the Grid corpus in the known configuration scenario.

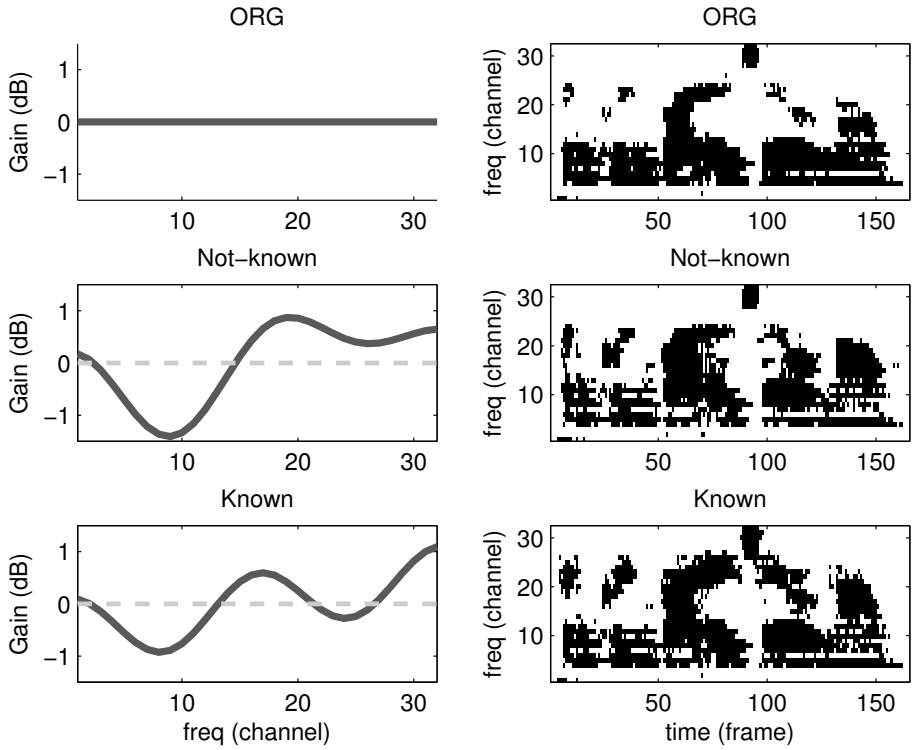

Figure 6.6: The average relative spectra of a *female* talker of roughly 20 utterance (left panels) and example masks (right panels) of a target utterance 'bin green by a zero now' for known and not-known talker scenarios for DIS-OPT compared to ORG in SSN at 0 dB SNR: (i) *Not-known* denotes using a speaker-independent (SI) acoustic model in the optimisation process; and (ii) *Known* denotes using a speaker-dependent (SD) acoustic model in the optimisation process, covering the frequency range between between 50 and 8000 Hz. Note, the female talker is (id = 11).

6.5.3 Statistical analysis of known talker scenario across Grid corpus

For some applications where the talker is known, the use of speaker-dependant (SD) models, *i.e.,* when the acoustic models are trained on talker under test, may provide a good opportunity to improve intelligibility. This is supported by the idea that in noisy environments some talkers are consistently more intelligible than others Barker and Cooke (2007). Earlier studies found that the intelligibility of the same speech generated by different talkers is more likely to be different even if speech was generated under ideal setting scenarios of speaking and listening, *e.g.,* Bond and Moore (1994); Hood and Poole (1980). Additionally, Barker and Cooke (2007) found that a high degree of

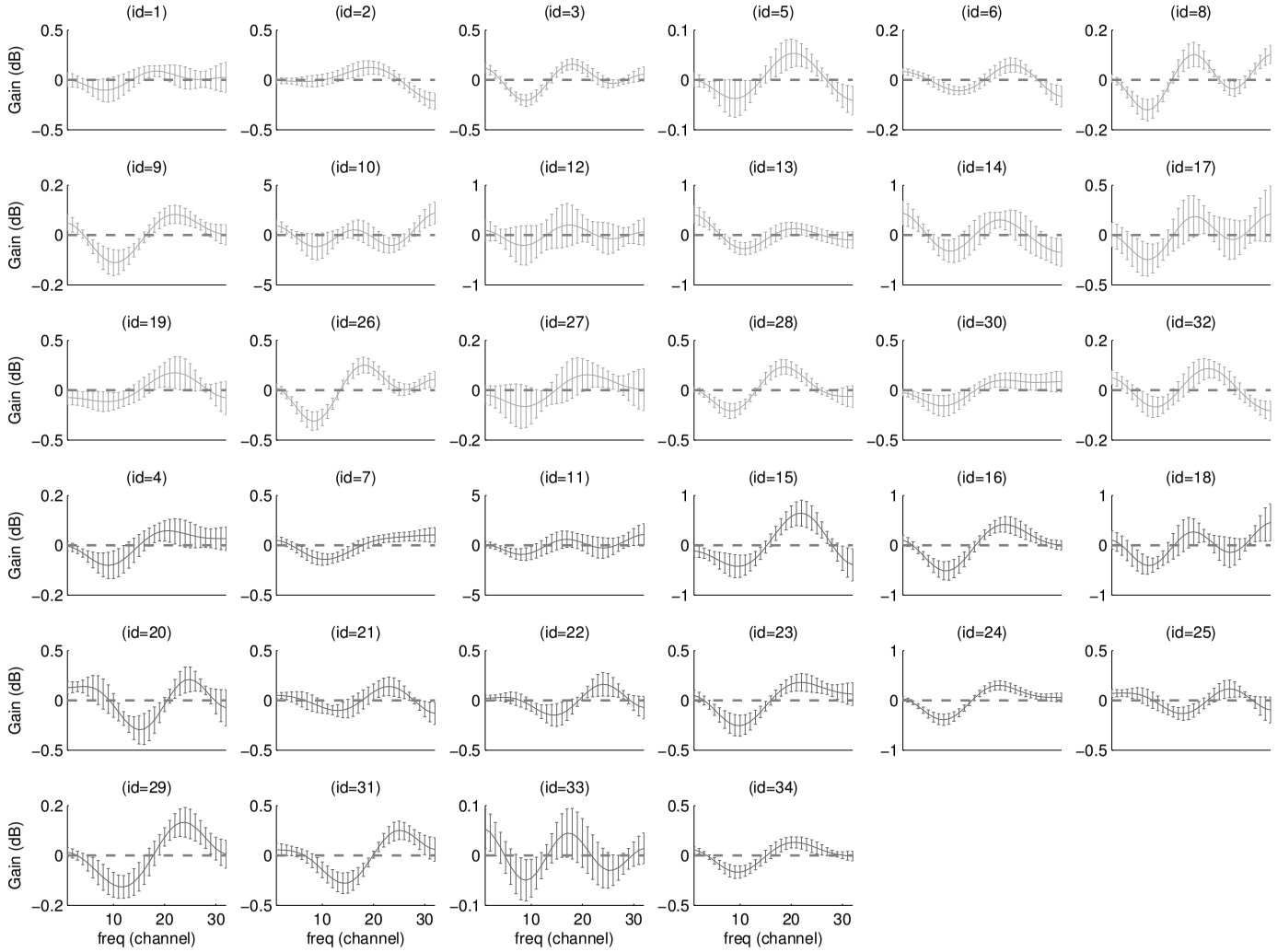

Figure 6.7: An illustration of the average relative spectra in (dB) with standard errors of modified speech in SSN across talkers in the Grid corpus, at a sentence level and averaged across a set of sentences per talker covering the frequency range between between 50 and 8000 Hz. The first three rows (*light* grey) represent **male** talkers and the last three rows (*dark* grey) denote **female** talkers. Note that the scale is different.

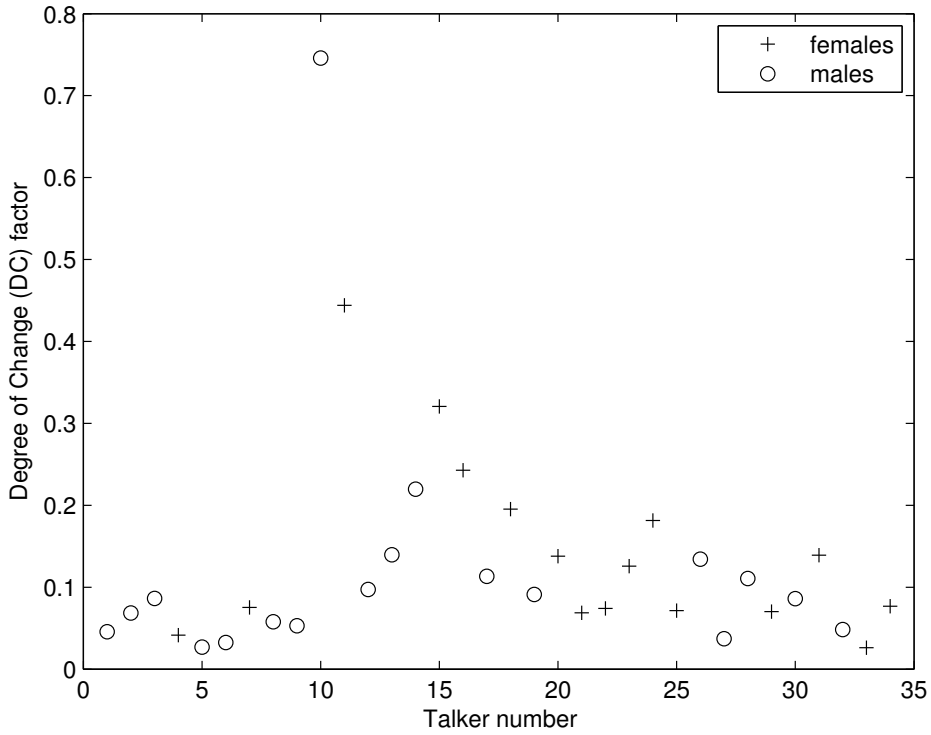

Figure 6.8: The degree of change (DC) factor of the TVDIS modified speech across male and female talkers in the Grid corpus. The DC computed as the absolute average of the average relative spectra for each individual talker separately for all talker in the Grid corpus.

variability in inter-talker intelligibility persists even when the SNR is the same across talkers.

The analysis is conducted to further examine the *(known)* talker scenario, reported in the proceeding experiment, across the 34 talkers in the Grid corpus. Precisely, we investigate whether there are some speaker who need to be more enhanced than others the experiment studies the impact of talker individual differences on generating the average relative spectra in SSN masker and hence attempting to find a similar trend among talker from the same gender.

We now aim to see the amount of change for each individual talker and relate the findings to the study by Barker and Cooke (2007) in terms of whether less hyper-articulated talkers, *i.e.,* their intrinsically intelligibility in noise is low, could benefit from this type of modification. Figure 6.8 demonstrates the DC factor of the TVDIS modified speech across male and female talkers in the Grid corpus. It can be seen that the amount of

gain is quite small for male and females talker and that the two talkers were benefited more which are talkers number 10 and 11, male and female, respectively.

Figure 6.7 shows the average relative spectra in (dB) with standard errors of DIS-OPT modified speech in SSN across talkers in the the Grid corpus, at a sentence level and averaged across a set of sentences per talker covering the frequency range between between 50 and 8000 Hz. The first three rows (*light* grey) represent **male** talkers and the last three rows (*dark* grey) denote **female** talkers. The results show that a similar pattern of the average shape between the talkers was found where the energy is reallocated from the either lower, middle or both to the higher frequency channels apart from talkers number 2, 5, 6, 13, 14, 25, and 32, respectively, in which some energy is taken from the higher channels to boost the middle frequency channels.

Comparing the average relative spectra of male speakers with the female speakers demonstrates that there is not an explicit difference among them. Nevertheless, a closer look at the average shape of the male talkers, the energy seems to centre in the middle frequency channels, while the higher frequency in more dominant in the female takers. This pattern is noticeable in the following female talker: 4, 7, 15, 16, 23, 24, 29 and 34. In general, these findings is in line with the results shown in the previous experiment 6.5.2 in terms of the balanced reallocation of energy between the channels, which is mostly centred in the middle and higher frequency channels. Thereby, boasting middle and higher channels is important in order to make speech more dominant in SSN maker.

The DIS-OPT works by reducing the discriminability between speech classes through making the important speech features more dominant, and hence eliminate the information masking. This is achieved through accounting for the individual difference of the talkers by using the SD acoustic model associated to the talker under test. On the other hand, the GP-OPT method used in the previous chapter works by reducing the energetic masking. In the following experiment, we evaluate the performance in terms of improving the intelligibility estimate of both methods using different objective intelligibility measures.

6.5.4 Performance analysis using macroscopic verses microscopic predictions of speech intelligibility

The impact of individual differences on the average relative spectra in the SSN masker was studied in the proceeding experiment. The aim of the experiment here is to objectively evaluate the effect of the DIS-based spectral modification system on speech

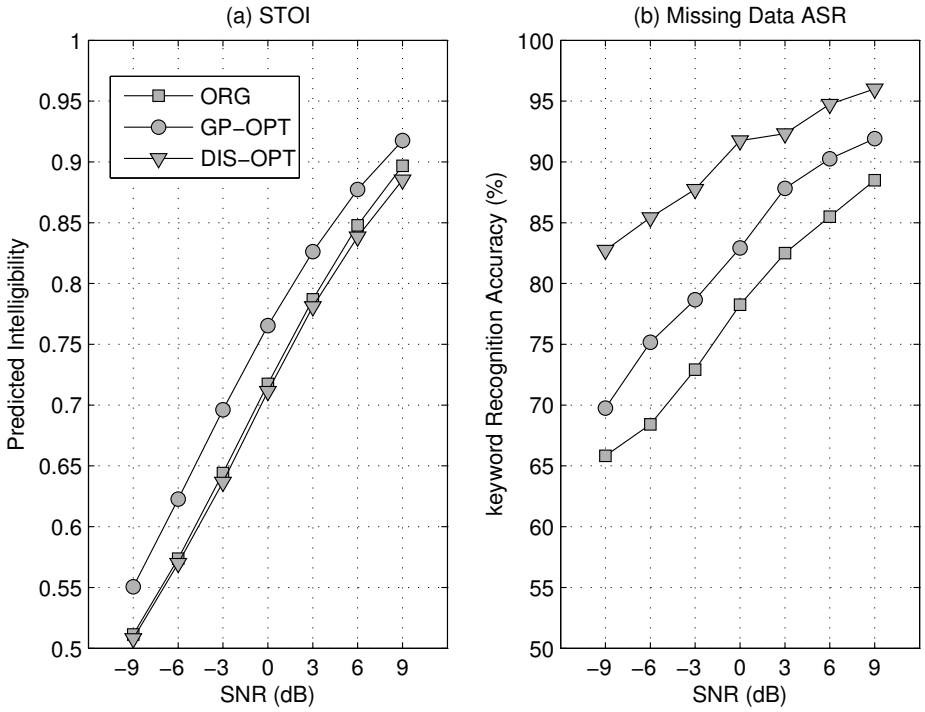

Figure 6.9: The predicted intelligibility using: (a) macroscopic - STOI - and (b) microscopic - missing data ASR - predictions of speech intelligibility in SSN at a range of SNRs for ORG, GP-OPT and DIS-OPT respectively.

Table 6.3: The predicted intelligibility scores using the Glimpse Proportion (GP) with standard deviation of error at a range of signal-to-noise-ratio (SNR) for different number of ORG, GP-OPT, and DIS-OPT speech conditions in SSN. Note that N was set to 4 in both GP-OPT, and DIS-OPT speech conditions.

	Method		
SNR (dB)	ORG $GP\ (\sigma_e)$	GP-OPT $GP\ (\sigma_e)$	DIS-OPT $GP\ (\sigma_e)$
-9	9.8 (0.05)	11.4 (0.05)	9.1 (0.07)
-6	13.2 (0.08)	16.8 (0.07)	13.0 (0.09)
-3	18.2 (0.11)	23.2 (0.10)	17.6 (0.13)
0	23.1 (0.14)	30.0 (0.13)	22.9 (0.18)
3	28.3 (0.16)	37.4 (0.16)	28.9 (0.21)
6	33.9 (0.18)	44.8 (0.19)	35.1 (0.26)
9	39.8 (0.18)	52.2 (0.21)	41.6 (0.28)
Average Score	23.8 (0.13)	30.8 (0.13)	24.0 (0.17)

intelligibility using macroscopic and microscopic predictions of speech intelligibility. For more information about the differences between the macroscopic and microscopic predictions measures and their computation procedures see Chapter 2.

In order to evaluate the intelligibility of the DIS-OPT modified speech macroscopically, we use a correction-based, *i.e.,* the STOI measure Taal et al. (2011), and SNR-based measures, *i.e.,* the GP measure Cooke (2006), with standard deviation of errors at a range of SNRs in SSN noise type. For the microscopic evaluation, we use the missing data based ASR Barker et al. (2005) and thus the keywords is considered as an estimate of intelligibility.

Figure 6.9 - case (a) - and Table 6.3 show the intelligibility estimate using macroscopic measures which are STOI and GP, respectively. The predictions clearly indicate that the GP-OPT outperform both the ORG and DIS-OPT, and that the DIS-OPT has a similar behaviour as the ORG case. For instance, the average GP score of the ORG across the SNRs is 23.8% which is roughly equal to the average GP of the DIS-OPT score, 24.0%. These findings may be better interpreted by the aim behind each method. That is, the ultimate goal of the GP-OPT is to reduce the energetic masking by increasing the number of of glimpses whereas the DIS-OPT does not attempt to increase the number of glimpses instead it selects the most informative glimpses.

However, the the microscopic evaluation using the missing data ASR, illustrated in Figure 6.9 - case (b), show a high degree of improvement of the DIS-OPT compared to the ORG and GP-OPT. In particular, we can compare the intelligibilities predicted by the microscopic model by looking at the percentage keywords that are predicted to be recognised correctly. For the ORG speech the intelligibilities range from 66% to 89%. The GP-OPT approach boosts these scores from 70% to 92%. For the DIS-OPT approach the scores are increased from 83% to 96%. Note, this HMM-based microscopic model has been previously evaluated in stationary noise masking conditions and has been shown to be a good predictor of listener performance.

6.6 Human Listening Experiment

In this section we present the results of a formal listening test. We contrast a number of speech types generated using different near-end intelligibility enhancement algorithms and also compare the results to the state-of-the-art and original speech. These speech types are listed in Table 6.1 which includes: ORG, GP-OPT, DIS-OPT, and SS. The outcome of this section has been published in Al Dabel and Barker (2015).

Table 6.4: The p-values for comparing intelligibility rates between techniques across SNRs levels.

	ORG	**GP-OPT**	**DIS-OPT**	**SS**
ORG	-	0.024	0.008	0.020
GP-OPT	0.024	-	0.060	0.015
DIS-OPT	0.008	0.060	-	0.015
SS	0.020	0.015	0.0156	-

6.6.1 Experimental Design

This section re-illustrates the same experimental design represented in the previous chapter in Section 5.6. As stated before, twenty normal-hearing subjects whose age ranged from 18 to 30 years participated in the listening tests. The subjects were required to be native English speakers, with no history of speech and/or language dis-orders. All were paid for their participation. Ethics permission was obtained.

The subjective evaluation were preformed in 5 noise conditions using a total of $13,600$ stimuli (4 algorithms x 680 sentences (34 talkers x 20 utterances) x 5 conditions) divided into independent blocks of 136. The independent block was drawn at random, without replacement in which a single subject would hear 34 sentences from each entry into the 5 blocks (34 x 5 = 170 sentences in total). The subjects were assigned into blocks in which; (i) each subject heard one block of 136 (34 utterances x 4 algorithms) sentences in each of the 5 noise conditions; (ii) no subject heard the same sentence twice; and finally (iii) each noise condition was heard by the same number of subjects. Subjects were tested individually in an acoustically-isolated booth. Stimuli were presented once only. The task was to identify the **letter** and **digit** spoken and type the heard keywords. Once a participant had typed a response, the subsequent stimulus was presented automatically. Null responses were not permitted. The test was completed on average in 45 minutes.

6.6.2 Results and discussion

Figure 6.10 shows the actual recognition rates together with standard errors averaged across listeners for the four speech types as a function of SNR. The reported scores were computed as the average of percentage of correctly identified letters and digits. From this data, it is apparent that the performance of GP-OPT and SS is doing similarly well across SNRs.

A two-way repeated measures ANOVA with two within-subjects factors (SNR level

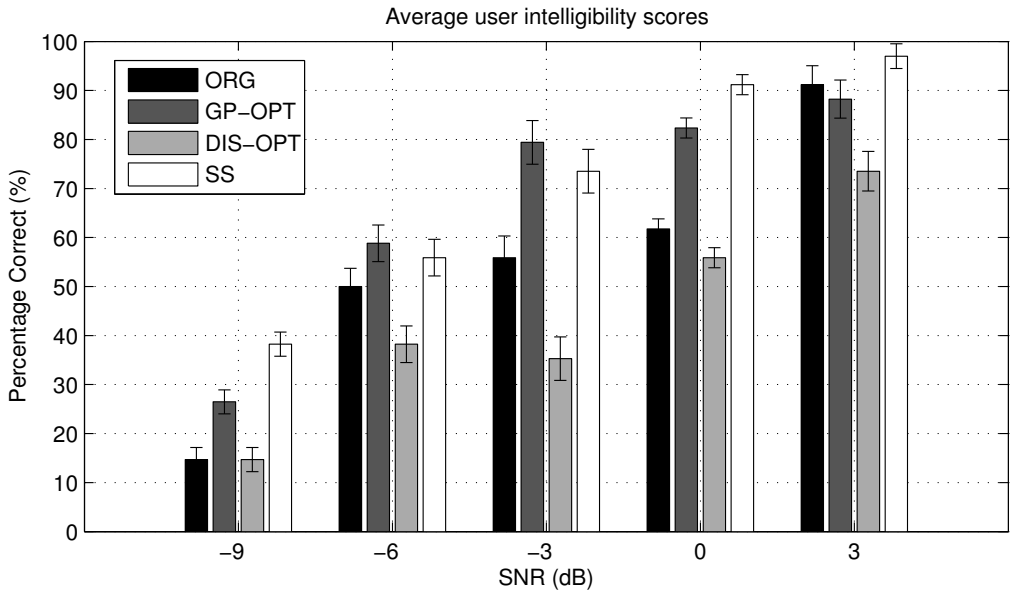

Figure 6.10: Percentage of correct identifications of both letter and digit in each speech type with the standard errors in the speech-shaped noise. The *black* shows performance in the ORG condition as a function of the signal-to-noise-ratio (SNR) used to produce the stimulus. The *dark grey* in the figure shows performance in the GP-OPT condition as a function of the SNR used to produce the stimulus. The *light grey* in the figure shows performance in the DIS-OPT condition as a function of the SNR used to produce the stimulus. The *white* shows performance in the SS condition as a function of the SNR used to produce the stimulus. Note that all were presented at the same overall RMS level.

and algorithm type) revealed that a gradual positive impact of SNR level ($F(4,12) = 48.67$, $p < 0.05$), and a significant effect of algorithm type ($F(3,12) = 16.58$, $p < 0.05$). Our primary purpose is to understand if there is an interaction between these two factors on the overall intelligibility. There was a significant difference between the ORG and the remaining entries. When comparing the performance of ORG against both GP-OPT and SS, for instance, it can be seen that the overall intelligibility rate for higher and lower SNRs levels was a nearly equivalent compared to the ORG with a difference equivalent to about 10 % of performance. However, the performance of DIS-OPT is comparable to the ORG across SNRs except at -3 dB.

A further statistical analysis was carried out using a pairwise comparison analysis. Mean intelligibilities are 54.7 %, 67.0 %, 43.5 % and 71.7 % for ORG, GP-OPT, DIS-OPT and SS across SNRs levels, respectively. A major difference can be seen between GP-OPT and DIS-OPT with difference in mean of 23.5 % and between SS and DIS-OPT of 28.23 %. The p-values can be found in Table 6.4.

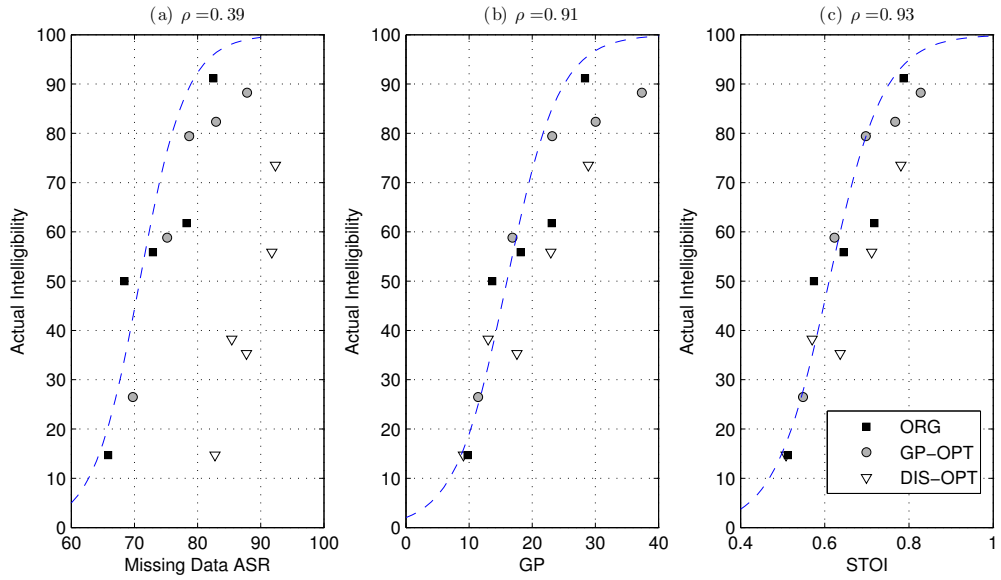

Figure 6.11: Scatter plots of (a) missing data ASR predictions versus actual intelligibility scores, (b) GP predictions versus actual intelligibility scores, and (c) STOI predictions versus actual intelligibility scores.

It is interesting to test whether these listening tests results could have been correctly predicted by recent objective measures of intelligibility. Figure 6.11 shows the results of scatter plots of (a) missing data ASR predictions versus actual intelligibility scores, (b) GP predictions versus actual intelligibility scores, and (c) STOI predictions versus actual intelligibility scores at a range of SNRs. It can be seen that the GP and STOI measures shown in Figure 6.11 have a high level of agreement with the actual listening results compared to missing data ASR. In summary, we have evaluated a number of different motivated near-end intelligibility enhancement systems aiming at better recognising the keywords of a target utterance in the presence of SSN. In particular, we compare the performance of both a reference in an open-loop system and developed spectral shaping in a close-loop feedback system and see whether using a priori knowledge of sound sources in the mixture and/or pre-trained speech models would result in a high intelligibility gain.

Our analysis shows that the spectral shaping, developed here, using a simple measure of energetic masking is more likely to improve speaker intelligibility on the listening task and obtained a significant improvement over the baseline performance across all SNR conditions. However, the developed spectral shaping using a more complicated measure of decoding the target utterance result in a poor correlation with the intelligibility both

across SNRs and within a single SNR. The most striking result to emerge from the data is that the priori knowledge of SSN noise might not correspond to the intelligibility improvement compared with the reference spectral shaping. This claim has been made in Cooke et al. (2013a).

The DIS-OPT system is primarily based on a measure of decoding using the DIS model. The DIS accounts for the entire target utterance, and not specific to parts of the utterance to be enhanced although the motivation behind this measure was to discriminate between the correct class with the class that is most acoustically similar. The nature of listening task, however, was to identify the letters and digits in the spoken utterance in noise. Hence, the development of the system and the listening task might be not compatible and that might justify the poor performance of the enhancer despite the extra-embedded knowledge.

For the developed spectral shaping to tackle advance conditions, one could extend the algorithm by adding a time-varying modification. This extension might make an improvement in the intelligibility of speakers. One could further extend the DIS model by decoding the parts of utterance that is believed to be important for intelligibility or reducing the confusing. In the following chapter, we will explore these opportunities to further advance the system.

6.7 General Discussion

Increasing the contrast between the acoustically-similar classes of speech was the aim of the near-end pre-enhancement system built in this chapter. It has been based on using a pre-trained speech model of a single speaker. We hypothesise that a pre-enhancement system is more likely to benefit from the speaker-related extra knowledge once it is embedded within the optimisation. The novel contribution in this chapter is the discriminative microscopic intelligibility measure which has been developed and used within the pre-enhancement framework.

On average, we found a bigger impact in the observed average relative spectra for the discriminative based pre-enhancement system using both SI and SD acoustic models over the ORG, demonstrating a significant trend of moving the energy from lower to the middle frequencies (see Figure 6.4). Individual speakers did benefit slightly from the proposed system across the Grid corpus, none shown a significant change in the observed average relative spectra when compared to the ORG (as shown in Figure 6.7). By assessing the system objectively using macroscopic intelligibility (*i.e.,* STOI and

GP), we found that the discriminative based pre-enhancement system gave a roughly same intelligibility estimate as the ORG. This contrasts with the intelligibility estimate obtained based on the microscopic intelligibility (*i.e.,* missing data ASR) in which the proposed system outperform the ORG across the SNRs (as demonstrated in Figure 6.9). Based on humans evaluation, The DIS-OPT system differed significantly from the SS and the GP-OPT systems and it was even preforming poorly compared to the ORG speech. The listening test performance showed no improving despite the fact that the microscopic predictions showed an increase in the intelligibility estimate.

The discriminative microscopic intelligibility model sheds light on the important acoustic features that are believed to underlie the intelligibility of speech in noise by focusing on phonemes classes. In contrast, the GP (Cooke, 2006) measure used in the previous chapter has focused on the unmasked acoustic features of speech without paying attention to the relative perceptual importance of these speech features. The DIS is the ratio of probability between the correct transcription and the most probable incorrect transcription (as a single candidate). This begs the question of whether this measure differs from a recently proposed measure by Petkov et al. (2013). Petkov et al. (2013) have used a conventional ASR to compute an objective measure in which no explicit alternative transcriptions of noisy speech was considered at all, constructing only the probability of the correct transcription. Furthermore, the DIS is based on the theory of microscopic intelligibility by employing the missing data ASR. It also provide the opportunity to be a speaker-dependent based measure through training models unique to the chosen speaker.

As discussed in Section 5.7, spectral modification applied in the literature is usually the outcome of an optimisation based closed-loop system motivated by the idea of reducing the amount of energetic masking (EM) in a known noise scenario (*e.g.,* (Petkov et al., 2013; Taal and Jensen, 2013; Tang and Cooke, 2012)). Similarly, in the DIS-OPT, the spectral shaping has been used as the the result of the optimisation based closed-loop framework in an attempt to reduce the confusability between the acoustically-similar speech classes in the known noise. Whereas eliminating the amount of EM was the main aim of previously proposed systems, the aim of the DIS-OPT system was to make the difference between the acoustically-similar speech classes more dominant, besides decreasing the amount of EM. This may introduce complexity to the system and thus may explain why the DIS-OPT system failed to improve the intelligibility in noise. The findings indicated that for the speaker tested, although global relative average spectra do appear to have some bearing on overall intelligibility, there is not an overall clear

speech benefit to improve the intelligibility at a sentence level.

There may be several factors associated with the poor performance of the DIS-OPT system including: the implementation procedure, the computational complexity, the optimisation settings, and the spectral-only modification method. In the DIS-OPT, we searched for a set of cepstral coefficients that defined the optimal shaping at an utterance-level. The optimisation therefore is implemented at an utterance-level. To overcome these shortcomings, we need a way to implement optimisation at a segmental-level instead of the utterance-level. It is likely that the proposed DIS measure is more applicable in a time-varying scenario rather than the spectral-only scenario.

6.8 Chapter Summary

This chapter has proposed a novel technique for optimising the spectral shaping applied by a pre-enhancement system. The approach seeks to maximise intelligibility with respect to a microscopic intelligibility model and is designed to be used in situations where both the speech and noise signal are known a priori and where a statistical model of the speech is available. Compared to approaches that only consider the degree of spectro-temporal masking, *e.g.,* GP-OPT), the new approach finds more varied utterance-dependent tunings of the shaping parameters. Evidence from a microscopic intelligibility model (similar to one that has been validated in the same noise conditions Barker and Cooke (2007)) suggest that the new approach produces a more intelligible result, whereas the listening test gives a contradictory results. Examples are provided on the authors' web site for readers to judge[2].

The experiments in this chapter have employed stationary noise and speech material that has come from a corpus that allows construction of very precise speaker-dependent models. These settings may be appropriate as a first approximation in many application settings, but we now wish to generalise the approach to handle more complex data. The work in the following chapter is examining dynamically varying shaping filters that will allow the enhancement to adapt to changing characteristics of the background.

[2] http://staffwww.dcs.shef.ac.uk/people/m.aldabel/

Chapter 7

Time-varying Spectral Modification

7.1 Introduction

In the previous chapters, we considered stationary maskers to develop our near-end intelligibility enhancement systems in order to improve speech intelligibility in noise. It is known that stationary maskers induce mainly a peripheral masking that happens when energy from the target speech and the masker overlaps both spectrally and temporally (Brungart, 2001). The stationary maskers, however, are not representative of the types of noise environment typically experienced in everyday situations. Thus, we aim to extend the developed near-end intelligibility enhancement so that it can be applied in situations where the masker is fluctuating. This is more typical of everyday listening situations.

In contrast to steady-state maskers that mainly yields energetic masking, fluctuating maskers such as competing speech and speech-modulated noise lead to both energetic and *informational* masking. Informational masking is any additional decrease in speech intelligibility once the energetic masking in the auditory periphery has been considered, readers are referred to Chapter 2 (Section 2.2.2) for a discussion about energetic and informational masking.

The goal of this chapter is to extend the spectral modification strategy, developed in Chapter 5, to accommodate the time-varying changes in the spectro-temporal domain, for a better intelligibility improvement in the presence of fluctuating masker. Therefore, a spectral shaping is defined on a frame-by-frame basis instead of defining a band-dependent spectral shaping. This stage of work consists of three steps: (i) we propose to segment the spectrogram using information extracted from the speech waveform, (ii) spectral shaping is extended into the time-varying so that the energy is distributed

spectrally and temporally, and (iii) the optimisation is performed locally on a segment-by-segment basis, instead of the entire spectrogram globally.

We also aim to validate the performance of the time-varying intelligibility enhancement systems. This will be achieved by comparing the proposed systems with the state-of-the art using human listening test in steady-state and fluctuating maskers.

This chapter is organised into seven sections. Sections 7.2 presents the development of a time-varying spectral modification strategy. Sections 7.3 describes a way to segment speech priori to applying the time-varying spectral shaping. In Section 7.4, we show how to apply the model and the modification in our closed-loop enhancement system. The closed-loop near-end intelligibly enhancement systems are then objectively and subjectively evaluated in Section 7.5 and 7.6, respectively. The chapter is finally ends with a general discussion in Section 7.7 and a summary in Section 7.8.

7.2 Time-varying Spectral Shaping

The spectral modification method, proposed in Chapter 5, works by defining a weight for each frequency band, and thus the spectral shaping was stationary across time-frame within each band. In this section, we develop a time-varying spectral modification in which a weight is defined for each individual frequency and time-frame.

The time-varying spectral shaping generates a weight for each time-frame t. The weight is controlled by the cepstral coefficients, $\mathbf{c}_t = [c_{1,t} \dots c_{n,t}]^T$, and is defined as;

$$S_c(t,f) = \sum_{n=0}^{N-1} (C_t \times c_{n,t}) \cos(\frac{\pi}{F}(n+\frac{1}{2})f), \tag{7.1}$$

where $t = 1, \dots, T$, and T is the total number of frames. Further, $f = 1, \dots, F$ and F is the number of frequency bands. Furthermore, the C_t is a constant to ensure that the location of the coefficients using a linear interpolation. In particular, the constant is defined as;

$$C_t = \begin{cases} 2(t_T - t)/T & \text{if} \quad t < t_c \\ 2(t - t_1)/T & \text{if} \quad t \geq t_c \end{cases}$$

in which t_1, t, t_c, and t_T are the first, current, centre, and last frames of the speech spectrum, respectively.

Computing the weight using the Equation 7.1 for $t = 1, \dots, T$ and $f = 1, \dots, F$

gives:

$$\mathcal{S}_c = \begin{bmatrix} S_c(1,1) & S_c(2,1) & \cdots & S_c(I,1) \\ S_c(1,2) & S_c(2,2) & \cdots & S_c(I,2) \\ \vdots & \vdots & \ddots & \vdots \\ S_c(1,F) & S_c(2,F) & \cdots & S_c(I,F) \end{bmatrix} \tag{7.2}$$

in which $\mathcal{S}_c$ denotes a weighting matrix of the same size $(T \times F)$.

Finally, we obtain the modified spectrogram $\hat{X}(t,f)$ by adding the spectro-temporal weight, $\mathcal{S}_c(t,f)$, to the original spectrogram $X(t,f)$;

$$X_c(t,f) = X(t,f) + \mathcal{S}_c(t,f) \tag{7.3}$$

where the energy remains unchanged before and after modification of spectrogram on a frame-by-frame basis.

7.3 Low Energy Based Segmentation Approach

In the previous section, we presented the time-varying spectral shaping strategy. We propose to apply this strategy on a segment-by-segment basis to reduce the effect of fluctuating masker. To do that, in this section we will define the segments first and then show how to apply the time-varying shaping on a segment-by-segment basis.

7.3.1 Defining Segmentation Boundaries

A first stage of the time-varying concept developed in this chapter, is to find a way to define the segments of the speech spectrogram, $X(t,f)$. We propose here to use low energy points of the clean original speech signal, $x(t)$.

To find the low energy points of the $x(t)$, a number of steps has to be taken. First, the temporal envelope is estimated by taking the magnitude of analytical signal corresponding to speech signal in time-domain, $e(t)$. Fast envelope fluctuations are reduced by smoothing the previous estimated signal using a first order lowpass filter. The smoothed temporal envelope, $\bar{e}(t)$, is then passed through a peaks and dips detection algorithm that requires a difference of at least 0.3 between a dip and its surrounding in order to declare it as a dip, and the same applied to the detected peaks. The detected dips are defined as low energy points. These points are then divided by the sampling frequency, f_s, and times by 100 in order to represent equivalent time-frames in the speech spec-

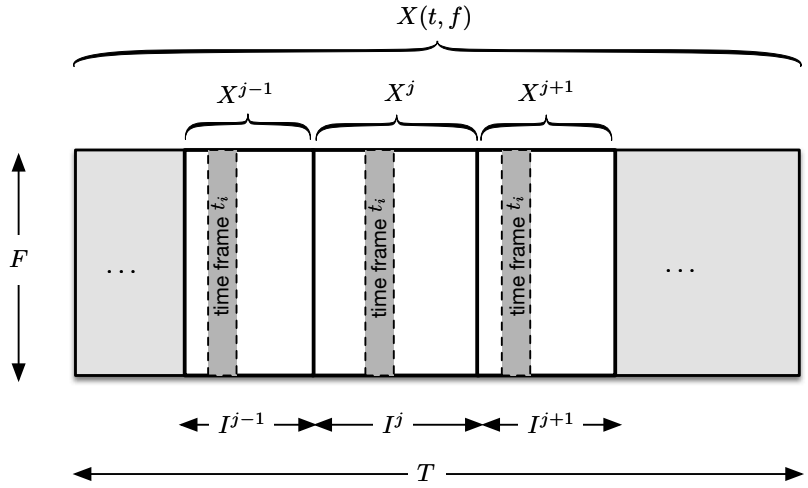

Figure 7.1: Schematic diagram of speech spectrogram segmentation into non-overlapping segments.

trogram. The resulting points are used to segment the clean original spectrogram into non-overlapping windows of different lengths.

In particular, we assume that an arbitrary speech spectrogram $X(t, f)$ of a clean speech signal is a $T \times F$ dimensional matrix (with T time frames and F frequency bands). The $X(t, f)$ can be expressed as a stacked combination of clean speech segments X^j, with $j = 1, \ldots, J$ denoting the segments index. These segments are spectrograms of size a $I^j \times F$ dimension extracted from the original speech spectrogram derived by aligning the time-frames associated with the low energy points to the $X(t, f)$, as illustrated in Figure 7.1. We thus write:

$$X(t, f) = [X^1(t_i^1, f) \ldots X^j(t_i^j, f) \ldots X^J(t_i^J, f)], \tag{7.4}$$

$$j = 1, \ldots, J, \quad \text{and} \quad t_i^j = t_1^j, \ldots, t_{I^j}^j \quad \text{where} \quad \sum_{j=1}^{J} I^j = T.$$

Similarly, we assume we have a $T \times F$ dimensional noise spectrogram $N(t, f)$. It can be represented by a stacked combination of K noise segments N^k, with $k = 1, \ldots, K$ being the noise segment index. It should be noted that the noise segments' lengths is the same as the clean speech segments' lengths in which $|K| = |J|$. Now we can write:

$$N(t, f) = [N^1(t_i^1, f) \ldots N^k(t_i^k, f) \ldots N^K(t_i^K, f)], \tag{7.5}$$

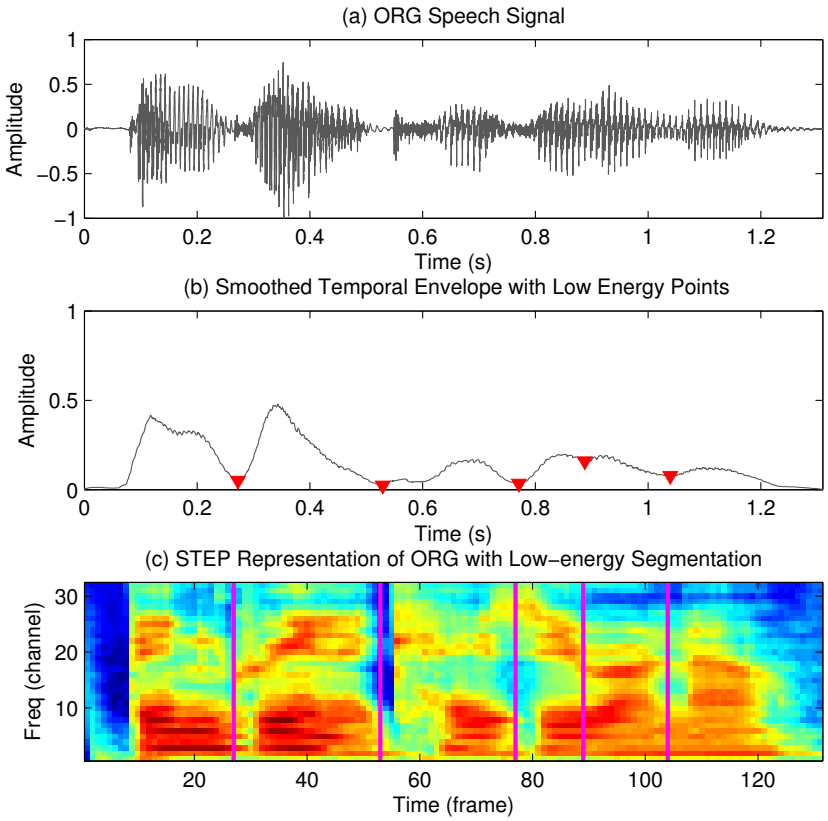

Figure 7.2: An illustration of the low energy based segmentation approach using a speech signal of a male talker speaking 'bin green at k zero now'. (a) The clean original speech signal. (b) The estimated smoothed temporal envelope of the speech signal with the detected low energy points. (c) The Spectro-Temporal Excitation Pattern (STEP) representation of the speech signal with the aligned low energy points. The STEP was produced by filtering the speech signal using a bank of 32 gammatone filters with centre frequencies spaced on an ERB-rate scale from 50 to 8000 Hz, then sampling the low-pass filtered Hilbert envelope at the output of each filter to 100 Hz and log-compressed.

$$k = 1, \ldots, K, \quad \text{and} \quad t_i^k = t_1^k, \ldots, t_{I^k}^k \quad \text{where} \quad \sum_{k=1}^{K} I^k = T.$$

The idea of the low energy based segmentation approach can be better explained by an example illustrated in Figure 7.2. The figure demonstrates the original speech signal of a male speaker from Grid corpus speaking 'bin green at k z now'. It also shows the smoothed estimated temporal envelope of the clean waveform alongside the detected low energy points. Finally, the speech spectrogram (*i.e.,* STEP representation which has been used throughout this thesis) is shown with the aligned frames represented the

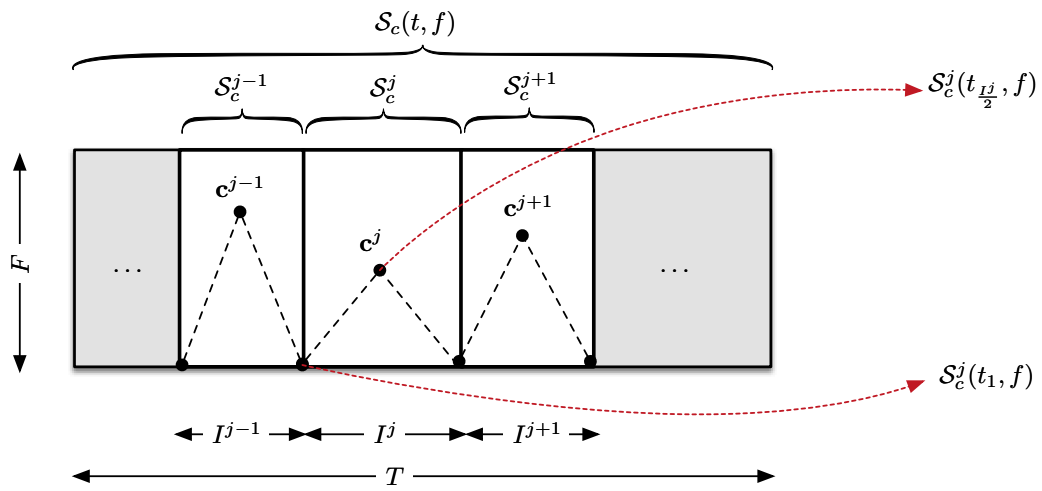

Figure 7.3: Schematic diagram of the local linear interpolation technique used in the time-varying spectral shaping to define the spectro-temporal weight on a segment-by-segment basis.

low energy regions of the speech signal.

In fact, it is difficult to decide the segmentation approach in which the clean spectrograms should be segmented. One, for instance, might define the word boundaries to be segment boundaries (*e.g.,* Petkov et al. (2013)), alternatively one might define the segmentation boundaries to lie at the vowel centres. Another feasible approach is to define the segments to have an equal fixed duration. A much more precise method of finding a segmentation approach is to detect the low energy points of the temporal envelope. This view is supported by the fact that the temporal envelope variations is reduced in the low-energy parts of speech signal (*e.g.,* nasals, onsets and offsets) (Allen, 2005a). Thus, the transition effects and artefacts, which are likely to be introduced later on due to potential discontinuity between speech parts, are prevented.

7.3.2 Applying Time-varying Spectral Shaping on a Segment-by-Segment Basis

Once the clean speech spectrogram is segmented, we will apply a time-varying spectral modification method on a segment-by-segment basis.

As explained in Section 7.2, the time-varying spectral shaping uses cepstral coefficients to generate a weight for each time-frame t_i of a given segment j, where the total number of segments are J. Similar to the Equation 7.1, the weight for a time-frame t_i,

which is controlled by the cepstral coefficients, $\mathbf{c}_{t_i}^j = [c_{1,t_i}^j \dots c_{n,t_i}^j]^T$, is shown as;

$$S_c^j(t_i, f) = \sum_{n=0}^{N-1} (C_{t_i}^j \times c_{n,t_i}^j) \cos(\frac{\pi}{F}(n + \frac{1}{2})f), \tag{7.6}$$

where $t_i = t_1, \dots, t_{I^j}$, and I^j is the total number of frames of the assigned segment j. In addition, $f = 1, \dots, F$ and F are the number of frequency bands. Furthermore, the $C_{t_i}^j$ is a constant to ensure that the location of the coefficients in j using linear interpolation in order to prevent discontinuity and the artefact that are more likely to introduced by moving from one segment to the other. This constant is defined as;

$$C_{t_i}^j = \begin{cases} (t_{I^j} - t_i)/2I^j & \text{if} \quad t_c \leq t_i \\ (t_i - t_1)/2I^j & \text{if} \quad t_c > t_i \end{cases}$$

in which t_1, t_i, t_c, and t_{I^j} are the first, current, centre, and last frames of the segment j respectively.

We now compute the weight using the Equation 7.6 for $i = 1, \dots, I^j$ and $f = 1, \dots, F$ gives:

$$\mathcal{S}_c^j = \begin{bmatrix} S_c^j(1,1) & S_c^j(2,1) & \cdots & S_c^j(I^j,1) \\ S_c^j(1,2) & S_c^j(2,2) & \cdots & S_c^j(I^j,2) \\ \vdots & \vdots & \ddots & \vdots \\ S_c^j(1,F) & S_c^j(2,F) & \cdots & S_c^j(I^j,F) \end{bmatrix} \tag{7.7}$$

in which $\mathcal{S}_c^j$ denotes a weighting matrix of the same size as segment j $(I^j \times F)$. A clarified diagram is shown in Figure 7.3.

Stack Equation 7.7 in column for each j results in the weighting matrix for $t = 1, \dots, T$, defined as follow;

$$\mathcal{S}_c(t, f) = [\mathcal{S}_c^1(t_i^1, f) \dots \mathcal{S}_c^j(t_i^j, f) \dots \mathcal{S}_c^J(t_i^J, f)], \tag{7.8}$$

Finally, we obtain the modified spectrogram $\hat{X}(t, f)$ by adding the spectro-temporal weight, $\mathcal{S}_c(t, f)$, to the original spectrogram $X(t, f)$;

$$X_c(t, f) = X(t, f) + \mathcal{S}_c(t, f) \tag{7.9}$$

where the energy remains unchanged before and after modification of spectrogram on a frame-by-frame basis.

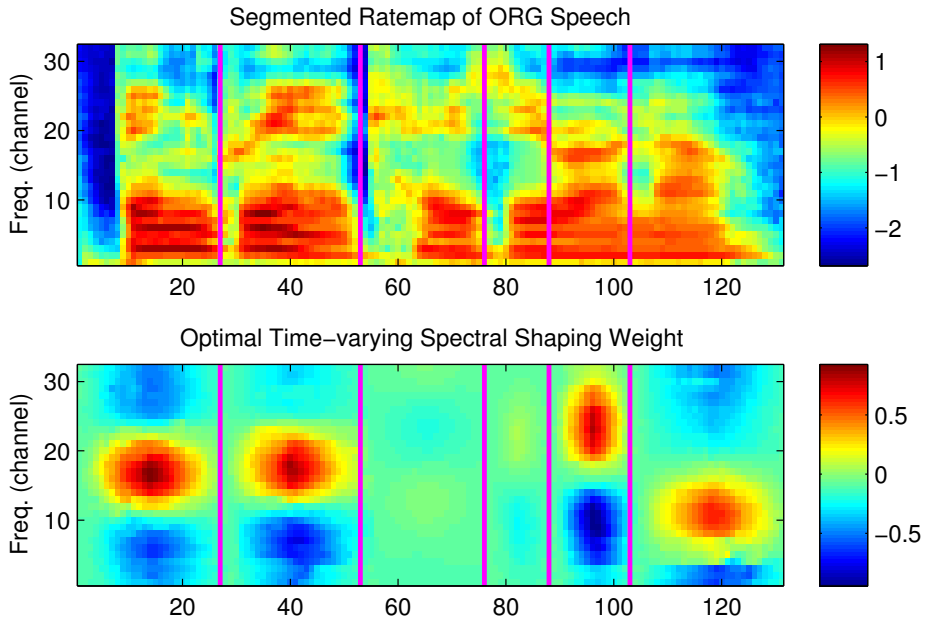

Figure 7.4: An illustration of time-varying spectral shaping approach using a speech signal of a male talker speaking 'bin green at k zero now'. The Spectro-Temporal Excitation Pattern (STEP) of the speech signal with the aligned low energy points (*in the upper panel*). The STEP was produced by filtering the speech signal using a bank of 32 gammatone filters with centre frequencies spaced on an ERB-rate scale from 50 to 8000 Hz, then sampling the low-pass filtered Hilbert envelope at the output of each filter to 100 Hz and log-compressed. The spectro-temporal weight using time-varying spectral shaping approach on a segment-by-segment basis (*in the lower panel*).

An example is illustrated in Figure 7.4. The figure shows the $X(t, f)$ of a male speaker from Grid corpus speaking 'bin green at k z now' with the aligned frames extracted using the low energy segmentation approach described in the former section. It also shows the example optimal time-varying spectral shaping weight applied on a segment-by-segment basis.

7.4 Optimising Intelligibility

In the previous sections, we described the components of the closed-loop intelligibility enhancement framework separately including: (i) the modification procedure using the time-varying spectral shaping (ii) the way in which the clean speech spectrogram is segmented, and then how to apply the time-varying spectral shaping on a segment-by-segment basis and finally (ii) the phoneme-level discriminative microscopic intelligibility

model. In this section, we bring all component together in order to optimise the intelligibility using the closed-loop enhancement framework.

Particularly, the aim here is to search for the optimal setting of system parameters based on maximising a chosen intelligibility measure on segment-by-segment basis. To do that, segment boundaries of clean speech spectrum need to be assigned first as described in Section 7.3. Once the segment has been assigned, the optimisation is then performed using the clean speech segment and noise segment (both of the same size) as inputs to the GP-based system. In the case of the phoneme-level DIS system, an additional input is required which is a model of the speaker to be enhanced.

Furthermore, the optimisation is preformed individually for each segment and thus the number of independent optimisation is $j = 1, \ldots, J$, is equivalent to the number of segments. Therefore, for a segment j, we define the problem formulation for GP-based closed-loop system (illustrated in Figure 7.5) as;

$$\hat{\mathbf{c}}^j = \underset{c}{\operatorname{argmax}}\, GP(X^j, N^j; c), \tag{7.10}$$

$$\text{where} \quad GP\left(X^j, N^j; c\right) = \frac{100}{I^j F} \sum_{t_i=0}^{I^j-1} \sum_{f=0}^{F-1} \mathcal{H}\left(\hat{X}_c^j(t_i, f) > \left(N^j(t_i, f) + \theta\right)\right). \tag{7.11}$$

where X^j and N^j are the speech and noise spectrum of segment j which denote the input requirements to the baseline system (illustrated in Figure 7.5). Further, I^j and F denote the numbers of time frames and frequency bands. The $\hat{X}_c^j(t_i, f)$ and $N^j(t_i, f)$ denote the STEP of the pre-enhanced speech and noise at time frame t_i and frequency band f of segment j, respectively. $\mathcal{H}(.)$ is the Heaviside step function counting the number of 'glimpses' which meet the local audibility criterion θ. The optimisation problem has been solved using the Nelder-Mead Direct Search method (described in Chapter 4 (Section 4.4.3).

Similarly, for the phoneme-level DIS-based closed-loop system (illustrated in Figure 7.6), we define the problem formulation as follows;

$$\hat{\mathbf{c}}^j = \underset{c}{\operatorname{argmax}}\, \mathcal{D}(X^j, N^j, \lambda; c), \tag{7.12}$$

using the DIS measure as described in Chapter 6:

$$\mathcal{D}(X^j, N^j, \lambda; c) = \sum_{t_i=1}^{I^j} \Big(\log P(\bar{X}_{t_i}|Q = Q_{t_i}^{(1)}) - \log P(\bar{X}_{t_i}|Q = Q_{t_i}^{(2)}) \Big) \tag{7.13}$$

$$= \sum_{t_i=1}^{I^j} \Big(\log P(\bar{X}_{t_i}|Q = Q_{t_i}^{(1)}) \Big) - \sum_{t_i=1}^{I^j} \Big(\log P(\bar{X}_{t_i}|Q = Q_{t_i}^{(2)}) \Big), \quad \bar{X} = \hat{X}_c^j, \tag{7.14}$$

in which X^j and N^j are the speech and noise spectrum of segment j, and λ is acoustic model of speech for the speaker to be enhanced. Both denote the input requirements to the phoneme-level DIS time-varying spectral shaping system. Regarding the $\mathcal{D}$, the left part of the equation denotes the likelihood of the correct states, $Q_{t_i}^{(1)}$, and the right part denotes the likelihood of the best scoring states, $Q_{t_i}^{(2)}$ with each time-frame, t_i, for segment j. The acoustic models that will be used in here in order to define the DIS are phoneme-based models. The phonemes models are represented as single-state for simplicity and to allow using the phoneme-level DIS on a segment-by-segment basis.

7.5 Objective Evaluation

Experiments were conducted to determine the impact of modifications methods on intelligibility in the presence of noise under energy and duration constraints.

First, Section 7.5.1 experimental setup and design used for the conducted experiments. Section 7.5.2 compared the performance of the the two modification systems described in this chapter. In Section 7.5.3, we compared the performance of the modification system developed in the previous chapter with the one developed in this chapter and relates the effect of gender differences. Further, we quantified the effect on intelligibility of modifications methods using objective intelligibility models namely STOI and GP in Section 7.5.4. Finally, the effect of applying the dynamic range compression on the intelligibility estimate on both original and modified speech were separately studied in Section 7.5.5 and 7.5.6, respectively.

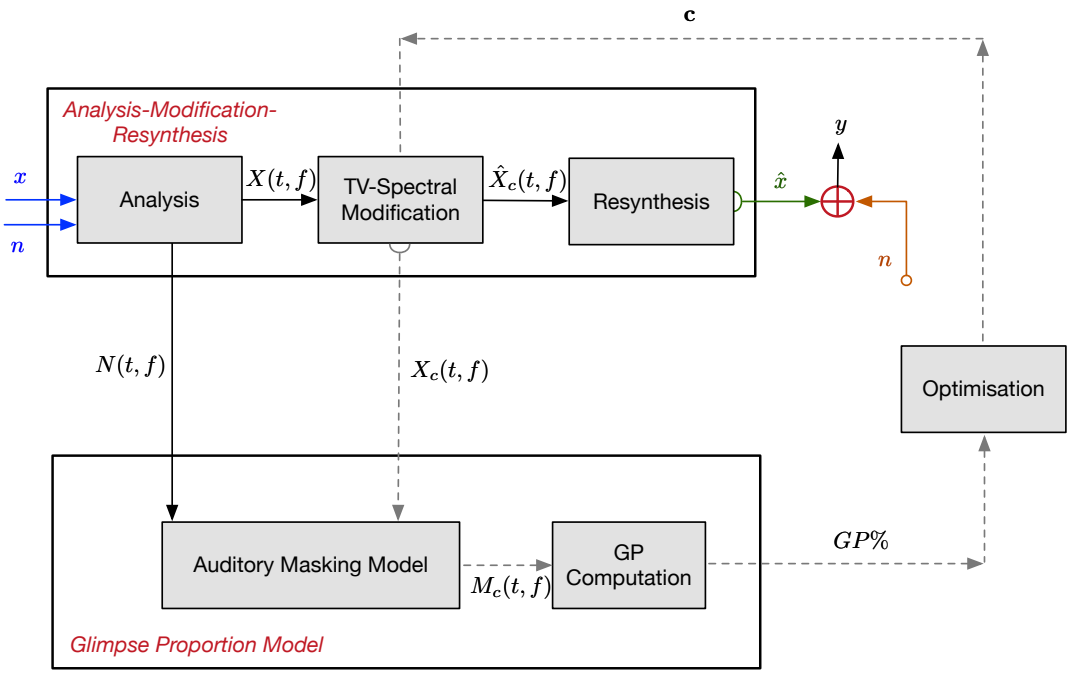

Figure 7.5: A schematic diagram of optimisation process for the GP-based time-varying spectral modification system, where x and n denote the input speech and noise signals. The $X(t, f)$ and $N(t, f)$ are the Spectro-Temporal Excitation Pattern (STEP) of x and n, respectively. The $X_c(t, f)$ denotes the modified and re-normalised STEP, used to compute the missing data mask $M_c(t, f)$. The $\mathbf{c}$ represents the optimal cepstral coefficients and GP_c is the Glimpse proportion metric. Finally, the optimal setting will result in the optimal STEP of speech $\hat{X}(t, f)$ in which it is resynthesised to represent the optimal enhanced and re-normalised signal $\hat{x}$. Solid arrows indicate fixed input to the optimisation process, whereas the grey dashed arrows indicate iterative process of optimisation. Note that we treat the $X(t, f)$ as one individual segment for presentation purposes.

7.5.1 Experimental setup

7.5.1.1 Speech types

The evaluation was conducted using the Grid corpus. The Grid corpus contains 34 native English speakers (18 male and 16 female), each of them speaking simple 6-word command sentences from a fixed grammar, recorded in a clean environment. There are 1000 utterances recorded from each speaker sampled at 25 kHz (refer to Section 5.5 for more details about the corpus). A subset data contains 680 utterances (20 per speaker) were used for the first evaluation.

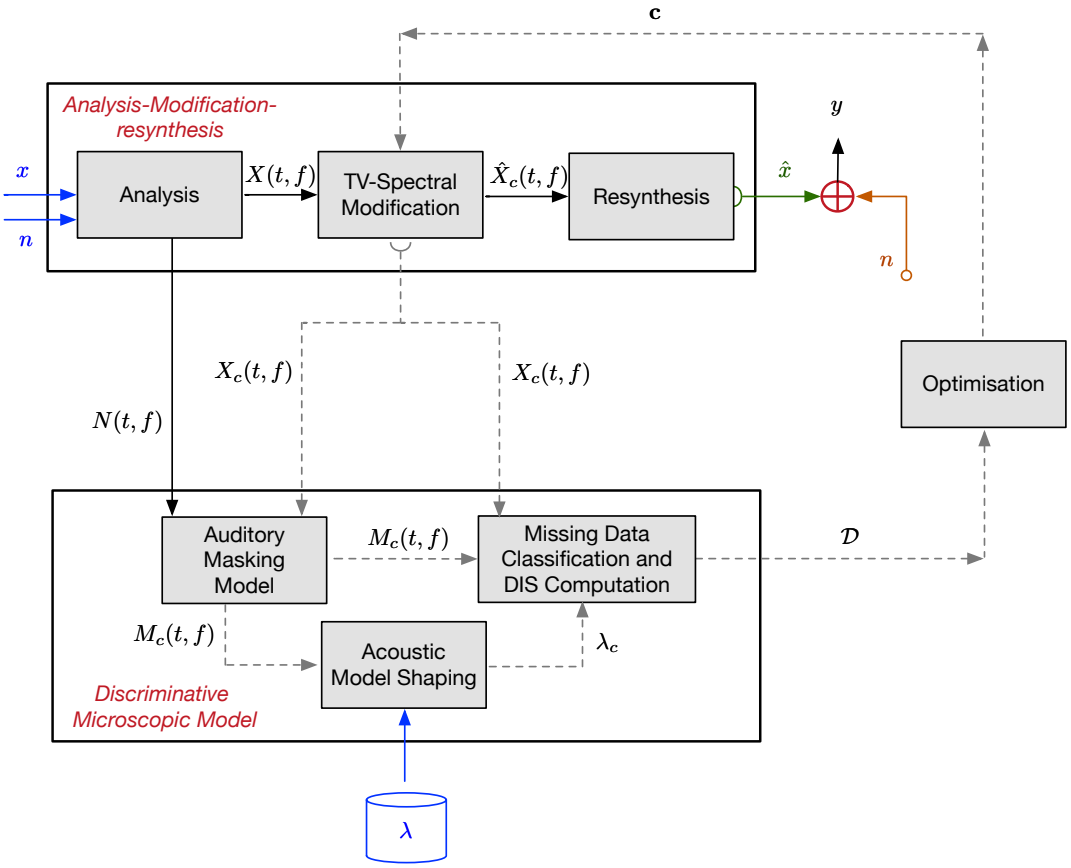

Figure 7.6: A schematic diagram of optimisation process for the phoneme-level DIS-based time-varying spectral modification system, where x and n denote the input speech and noise signals. The $X(t, f)$ and $N(t, f)$ are the Spectro-Temporal Excitation Pattern (STEP) of x and n, respectively. Further, λ and λ_c represents the speech model before and after shaping. The $X_c(t, f)$ denotes the modified and re-normalised STEP, used to compute the missing data mask $M_c(t, f)$. The $\mathbf{c}$ represents the optimal cepstral coefficients and $\mathcal{D}_c$ is the phoneme-level discriminative microscopic intelligibility model. Finally, the optimal setting will result in the optimal STEP of speech $\hat{X}(t, f)$ in which it is resynthesised to represent the optimal enhanced and re-normalised signal $\hat{x}$. Solid arrows indicate fixed input to the optimisation process, whereas the grey dashed arrows indicate iterative process of optimisation. Note that we treat the $X(t, f)$ as one individual segment for presentation purposes.

Table 7.1 lists the seven different speech enhancement systems we built for this evaluation alongside the required knowledge for modification. There was one unmodified natural type 'ORG' and the remaining are modified speech. The 'TVGP' and 'TVDIS' denote the time-varying spectral modification using either the GP measure, or the DIS

Table 7.1: Speech types used for the evaluation in this chapter.

Type	*System*	*Required Knowledge*
ORG	Original unmodified speech	-
ORG-DRC	ORG combined with dynamic range compression (DRC)	Speech signal
TVGP	GP-optimised speech using time-varying spectral modification (TV) technique	Speech and noise signals
TVDIS	DIS-optimised speech using TV technique	Speech and noise signals, and a model of clean speech
TVGP-DRC	TVGP-optimised speech combined with DRC	Speech and noise signals
TVDIS-DRC	TVDIS-optimised speech combined with DRC	Speech and noise signals and a model of clean speech
SSDRC	Spectral shaping combined with DRC based speech using Zorila et al. (2012)	Speech signal

model, respectively. Both are noise-dependent methods. We further processed the ORG, TVGP and TVDIS modified speech with the time-domain amplitude range compression strategy, refer to as DRC (a simplified DRC block diagram is illustrated in Figure 7.7). This results in the following modified speech: ORG-DRC, TVGP-DRC, and TVDIS-DRC. In order to compare the modified speech with the state-of-the-art, we further pre-process the clean Grid data with the spectral shaping and dynamic range compression system as described in Zorila et al. (2012), we refer to this system as SSDRC. The SSDRC system is noise-independent.

The TVGP and TVDIS were generated by first finding the optimal setting of the first four cepstral coefficients $\mathbf{c}$ (as shown in Figures 7.5 and 7.6 respectively). In particular, the speech was processed using a filterbank analysis-modification-synthesis framework. First, the speech signal is filtered using a bank of 32 gammatone filters with centre frequencies spread evenly on an equivalent rectangular bandwidth (ERB) scale between 50 and 8000 Hz with filter bandwidths matched to the ERB of human auditory filters. The instantaneous Hilbert envelope of each gammatone filter output is computed. This envelope is then smoothed by a first-order low-pass filter with an 8 ms time constant.

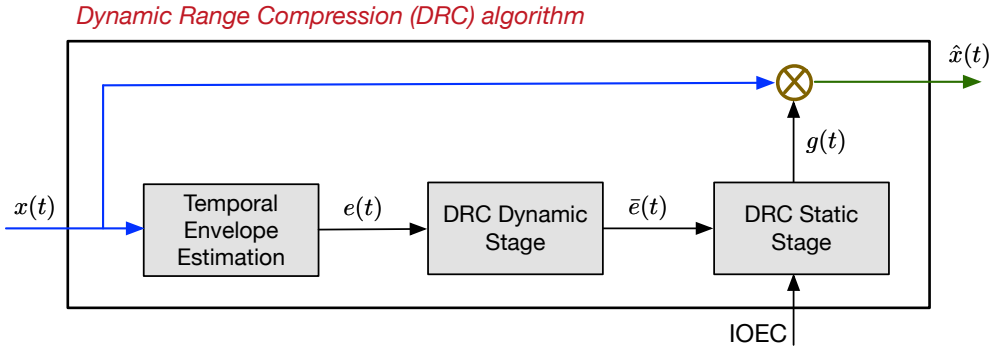

Figure 7.7: Block diagram of dynamic range compression (DRC) algorithm adapted from Zorila and Stylianou (2014). The $x(t)$ denotes the speech signal, and $e(t)$ and $\bar{e}(t)$ represent the temporal envelope of speech signal before and after applying DRC, respectively. Speech samples are re-scaled using time-varying gains $g(t)$ which are computed from the dynamically and statically compressed temporal envelope using a predefined input/output envelope characteristic (IOEC). The gains resulted from the previous operations are used to rescale the speech samples. Finally, energy-renormalisation is applied to ensure the same energy of the speech signal before and after DRC.

After that, the smoothed envelope is down-sampled to 100 Hz. After downsampling, the amplitude envelope is converted into the log-energy domain. The spectrum is then shaped by applying the optimal spectro-temporal weight, $\mathcal{S}_c$, to the gammatone filter outputs before resumming them to form the pre-enhanced signal. However, in order to ensure that the spectral shaping is smooth over frequency for each individual time-frame we consider only spectral shaping profiles that can be represented using the first N terms of a discrete cosine series. In this work N has been set to 4. Further, c_0 is arbitrarily fixed to 0 because it simply adds a constant gain factor across frequency that does not change the spectral shape. After scaling the filterbank outputs, resynthesis is employed to generate the spectrally shaped speech signal. Care needs to be taken when summing the bands to compensate for band-dependent phase delays introduced by the analysis. For details see Chapter 4. After resynthesis, the energy of pre-enhanced signal is scaled such that the global signal energy remains unchanged before and after spectral modification. The result is the enhanced signal, $\hat{x}$, that will be transmitted into the noisy environment.

7.5.1.2 Maskers

Speech enhancement systems were evaluated using both fluctuating and steady-state maskers at a range of SNRs. The steady-state masker was speech-shaped noise (SSN). This masker was generated by filtering white Gaussian noise through a 100-order all-pole filter, the long-term average spectrum of this noise was approximated to match that of the Grid speech material. Further, the fluctuating masker was a non-stationary **N**-talker babble modulated noise (BMN). This masker was generated by modulating SSN with the envelope of **N**-talker babble for various N. The envelope was calculated by convolving the absolute value of an **N**-talker babble signal with a 7.2 ms rectangular window. Babble was generated by summing utterances with equal rms energy from the Grid corpus. In this evaluation, **N** was fixed to 5.

7.5.1.3 Speech-noise mixtures

The SSN and BMN maskers were added separately to all speech types (see Table 7.1) at seven SNRs: 9, 6, 3, 0, −3, −6 and −9 dB. These SNRs were chosen to cover the full intelligibility range. The target utterances were mixed with the masker after the modification mechanism and energy renormalisation.

7.5.1.4 Acoustic speech models

To train acoustic speech models, a 17,000 utterance training set was provided containing 500 utterances from each of the 34 Grid speakers. We construct phoneme-level HMMs. The number of acoustic phoneme models K is 39. Each phone is modelled using a 3-state HMM with each state modelled as an 7-component diagonal covariance GMM. We first train a speaker-independent (SI) model from the full 17,000 utterances training set. Then we derived a speaker-dependent (SD) model for each of the 34 speakers by running further parameter re-estimations using just the target talker training data.

Before using the SD acoustic models in the modification system, we first pooled the 3-states for each SD phoneme-HMM model into a single state. We summed the priori of the 3-states and divided them by 3. Thus the number of GMM for each acoustic phoneme models is $3 \times 7 = 21$ GMMs.

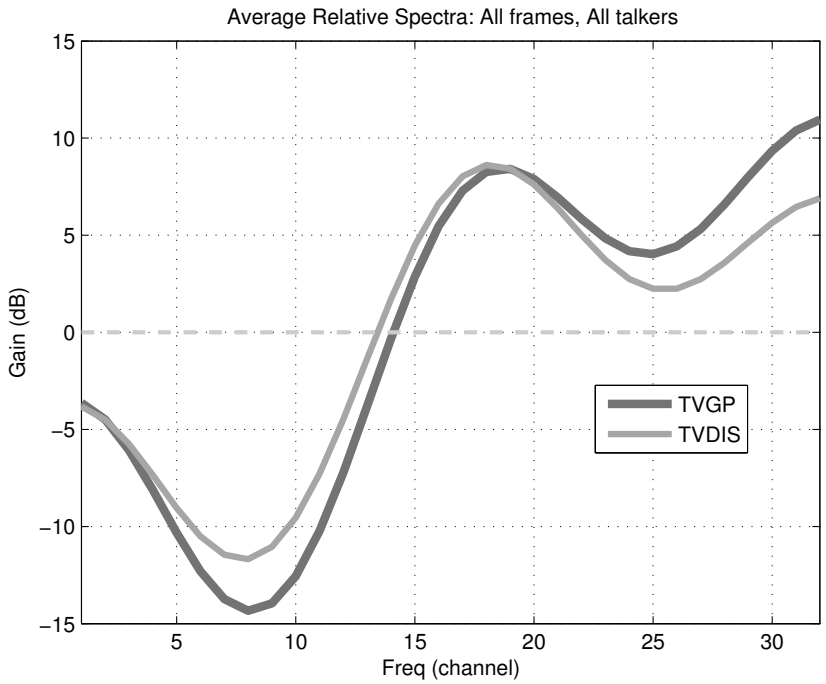

Figure 7.8: The average relative spectra for GP-based time-varying spectral optimised speech (TVGP) and DIS-based time-varying spectral optimised speech (TVDIS) averaged across all frames and all speakers in the Grid corpus in the babble-modulated masker, covering the frequency range between between 50 and 8000 Hz. The dashed line represents the original unmodified case (ORG).

7.5.2 Performance analysis of modification systems using average relative spectra

This section investigates the impact of the modification at the utterance level in terms of the average relative spectra. The 'relative spectra' is computed as the log differences between the average spectral envelopes of the ORG and modified speech calculated over all frames of the test data. The resulting relative spectra are then represented in (dB). We illustrate a broad analysis of a set of utterances that contains roughly 20 utterances per talker (20×34).

In particular, we calculated the the average relative spectra of the TVGP and TVDIS. Both TVGP and TVDIS modified speech were optimised for BMN masker. Figure 7.8 shows the overall pattern of the average relative spectra of the modified speech. It is clearly seen that the energy is reallocated from lower frequency channels (from channel 1 to channel 14) to the middle and higher frequency channels. Both TVGP and TVDIS have a similar trend in redistributing the energy with approximately the same amount

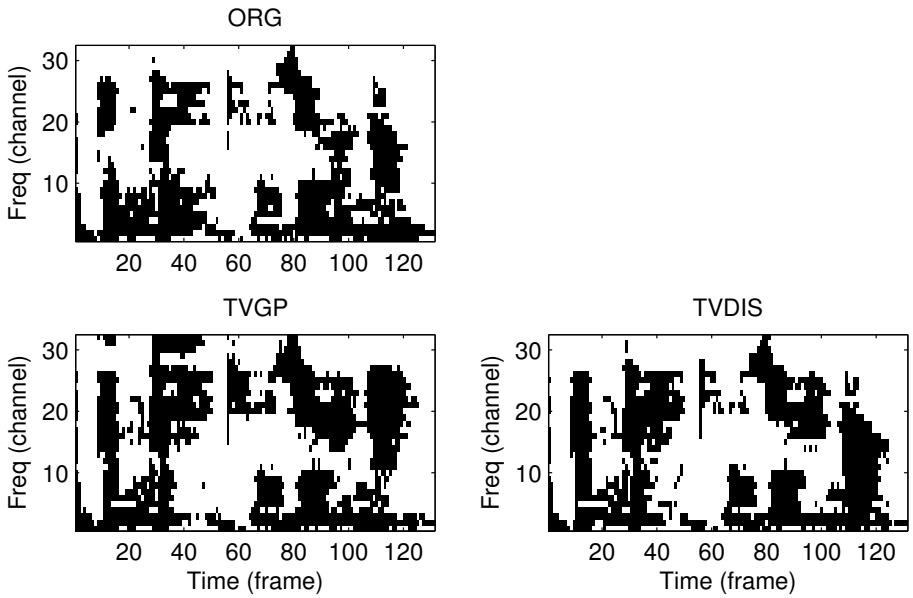

Figure 7.9: Example missing data masks of GP-based time-varying spectral optimised speech (TVGP) and DIS-based time-varying spectral optimised speech (TVDIS) compared to the original unmodified speech (ORG) of a male talker speaking 'bin green at k z now' in BMN at 0 dB SNR, using a bank of 32 gammatone filters with centre frequencies spaced on an ERB-rate scale from 50 to 8000 Hz. Note that both TVGP and TVDIS modified speech were optimised for *babble-modulated* masker.

of change, around 8 dB, in the middle frequency channels (from channel 14 to channel 23). However, the baseline system - TVGP - seems to reallocate slightly more energy from the lower to the higher frequency channels compared to the TVDIS system. This is due to the fact that the measure used during the optimisation is the GP metric which accounts for eliminating the energetic masking.

For a detailed comparison of the proposed TVGP and TVDIS methods in operation on a specific utterance, Figure 7.9 shows the missing data mask detected in the presence of BMN masker at 0 dB SNR for TVGP and TVDIS compared to the original unmodified speech ORG of a male talker speaking 'bin green at k z now', using a bank of 32 gammatone filters with centre frequencies spaced on an ERB-rate scale from 50 to 8000 Hz. We can see that more spectro-temporal elements of speech are visible at the middle frequency channels in the TVDIS missing data mask compared to the ORG mask. In the TVGP masks the amount of redistribution of the spectro-temporal speech elements is greater in comparison to the TVDIS mask. Precisely, more spectro-temporal speech elements are introduced at the middle and higher frequency channels. This is inline with the observation obtained from the average relative spectra.

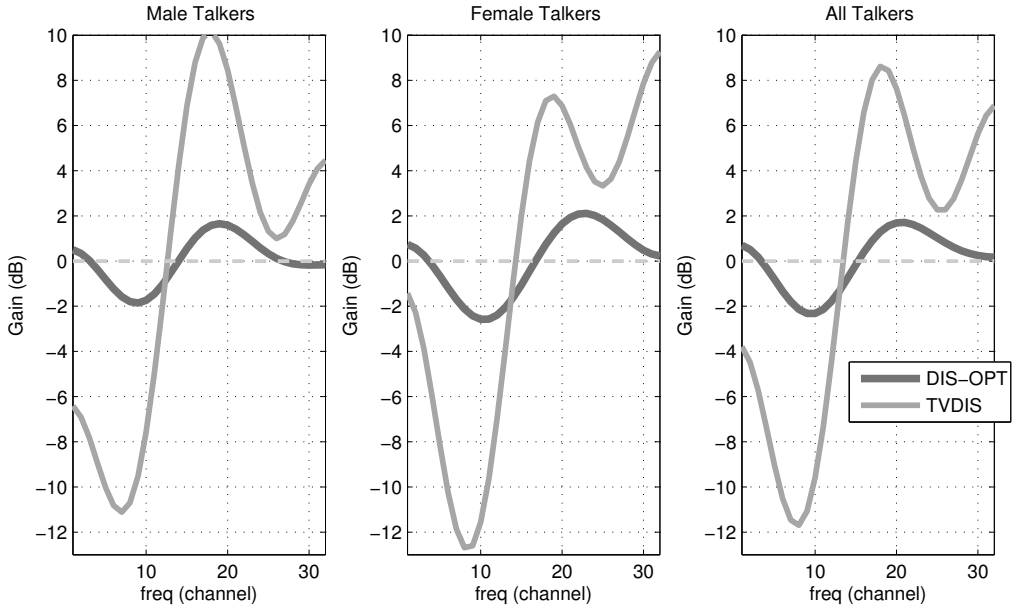

Figure 7.10: The average relative spectra for *male*, *female* and *all* talkers in the Grid corpus using the DIS-OPT and TVDIS modified speech optimised for speech-shaped noise, covering the frequency range between between 50 and 8000 Hz. The dashed line represents the original unmodified case (ORG). Note that the DIS-OPT was processed as described in the previous chapter.

7.5.3 Performance analysis of effect of gender differences on average spectral change

This section examines the behaviour of two modification systems on the male and female talkers in terms of average relative spectra. The two systems are: (i) the original version of the modification system developed in the previous chapter 'DIS-OPT' (see Figure 6.2) and (ii) the extended version of the modification system developed in this chapter 'TVDIS' (see Figure 7.6). We first compute the average the relative spectra for each individual talker separately of both DIS-OPT and TVDIS. We further average the data across male, female, and all talkers respectively. The modified speech in this analysis were optimised for SSN.

The question to be answered here is whether the spectral change in dB differs between males and females and how much does it differ by modified speech using both DIS-OPT and TVDIS. This is well demonstrated in Figure 7.10. The figure shows the average relative spectra for males, females, and all talkers of DIS-OPT modified speech compared to the TVDIS modified speech. It can be seen that for both types of modified speech

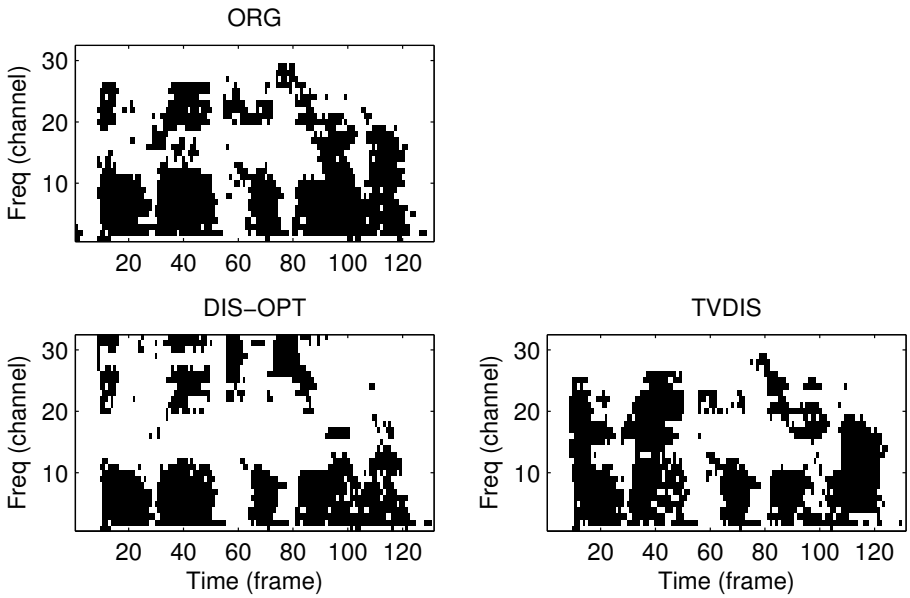

Figure 7.11: Example missing data masks of DIS-based spectral optimised speech (DIS-OPT) and DIS-based time-varying spectral optimised speech (TVDIS) compared to the original unmodified speech (ORG) of a male talker speaking 'bin green at k z now' in SSN at 0 dB SNR, using a bank of 32 gammatone filters with centre frequencies spaced on an ERB-rate scale from 50 to 8000 Hz. Note that both DIS-OPT and TVDIS modified speech were optimised for *speech-shaped* masker and the DIS-OPT was produced as described in the previous chapter.

the energy is concentrated at the middle frequency channels for the male talkers with the amount of increase is around 2 dB for DIS-OPT and around 10 dB for TVDIS in the middle frequency channels. In comparison, for female talkers, the energy in both types of modified speech seems to concentrate at the middle and higher frequency channels with the amount of increase is around 2 dB for DIS-OPT and around 6.5 dB and 8.5 dB for TVDIS in the middle and higher frequency channels, respectively.

In general, it can be hypothesised that the middle frequency channels are more essential for the intelligibility than the higher frequency channels for male talkers, whereas in the female talkers the reallocation of the energy differs. In the female talkers the middle and higher frequency channels are equally crucial for better intelligibility. On average, for all talkers the average relative spectra in the TVDIS gives roughly the same amount of energy to the middle and higher frequency channels.

As an example of both DIS-OPT and TVDIS methods, Figure 7.11 demonstrates the missing data mask detected in the presence of SSN masker at 0 dB SNR for DIS-OPT and TVDIS compared to the original unmodified speech ORG of a male talker speaking 'bin

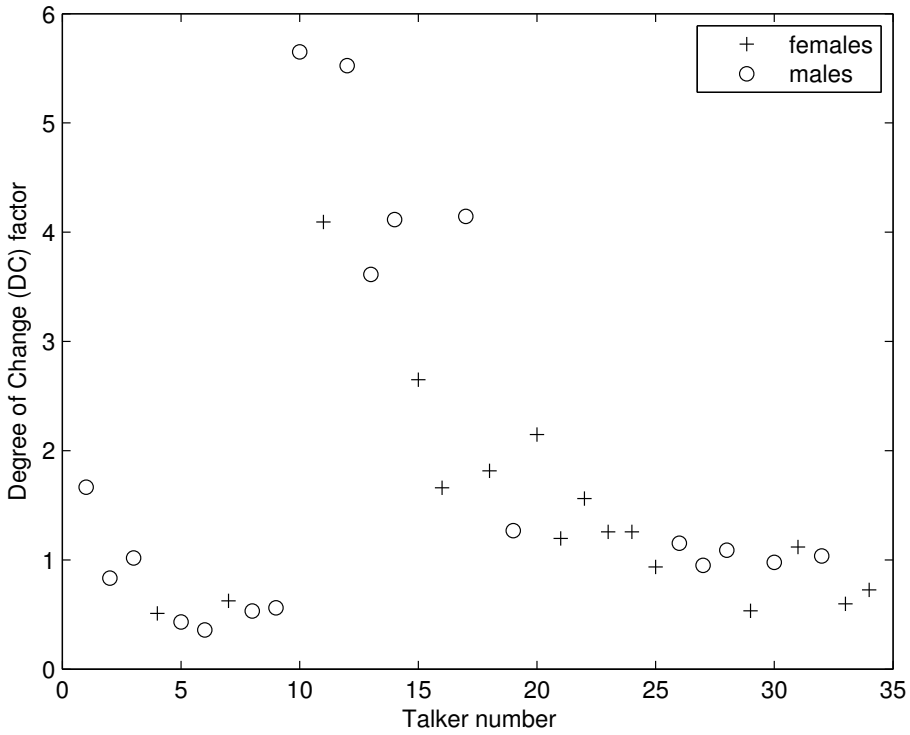

Figure 7.12: The degree of change (DC) factor of the TVDIS modified speech across male and female talkers in the Grid corpus. The DC computed as the absolute average of the average relative spectra for each individual talker separately for all talker in the Grid corpus.

green at k z now', using a bank of 32 gammatone filters with centre frequencies spaced on an ERB-rate scale from 50 to 8000 Hz. We can see that more spectro-temporal elements of speech are revealed in the middle frequency channels in the TVDIS missing data mask compared to the ORG mask. This is inline with the observation obtained from the average relative spectra.

To further investigate the amount of changes introduced among talkers, we define the degree of change (DC) factor. The DC factor is computed as the average absolute change of the average relative spectra. The amount of change for each individual talker has been measured and related to Barker and Cooke (2007).

Figure 7.12 shows the DC factor of the TVDIS modified speech across male and female talkers in the Grid corpus. The DC computed as the absolute average of the average relative spectra for each individual talker separately for all talker in the Grid corpus. It can be seen that on average the male talkers benefits more from the TVDIS

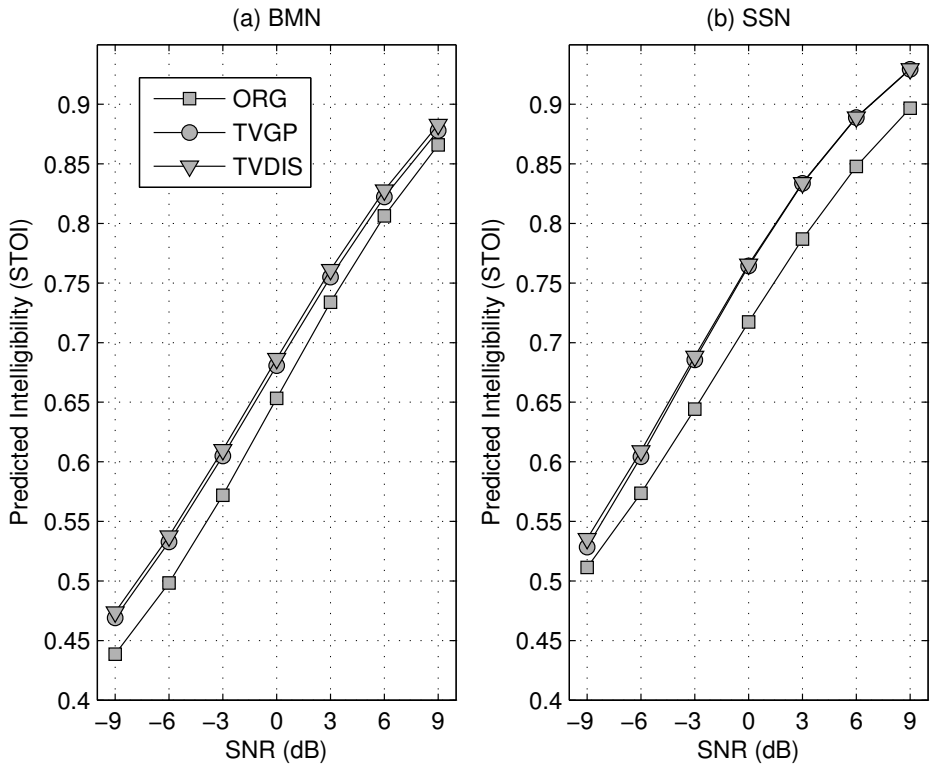

Figure 7.13: The predicted intelligibility scores using STOI at a range of SNRs for TVGP, and TVDIS based modified speech compared to ORG speech in BMN and SSN, respectively.

than the female talkers. This is may relate to the findings in Barker and Cooke (2007) where the intelligibility of male talkers in noise was found to be less than the intelligibility of female talkers in noise, and thus enhancing male talkers is better explained here. Now looking at the DC factor of individual talker of males and females and compare it to the intelligibility of that talker in noise using the study Barker and Cooke (2007). Therefore, it can be concluded that male speech is more likely to benefit from the enhancement system.

7.5.4 Performance analysis using objective intelligibility measures

The previous analysis examined the behaviour of modified speech introduced so far on the average relative spectra as well as relating the findings to measurements of the intelligibility of Grid talkers individually in noise.

Table 7.2: The predicted intelligibility scores using the GP at a range of SNRs for TVGP, and TVDIS modified speech in BMN and SSN.

	BMN			*SSN*		
SNR (dB)	ORG	TVGP	TVDIS	ORG	TVGP	TVDIS
-9	13.7	17.0	16.2	9.8	12.2	12.1
-6	18.8	23.0	22.1	13.2	18.0	17.5
-3	24.2	30.3	29.1	18.2	25.3	24.5
0	30.1	38.2	36.3	23.1	32.5	31.0
3	37.0	46.3	44.3	28.3	40.3	38.3
6	44.1	54.2	52.1	33.9	49.5	45.5
9	51.2	62.7	60.7	39.8	55.1	52.4
Average Score	31.5	39.2	37.7	23.8	33.7	31.5

Recent studies have focused on evaluating the impact of modification methods on improving the intelligibility of speech in fluctuating and steady-state noises in Cooke et al. (2013a,b). In this experiment, we quantify the intelligibility of the new modification methods in BMN and SSN maskers using objective intelligibility measures at a range of SNRs. The new modification methods includes TVGP and TVDIS modified speech (see Table 7.1). The intelligibility estimate is calculated using the Short-Term Objective Intelligibility (STOI) in Taal et al. (2011) measure and Glimpse Proportion (GP) metric in Cooke (2006).

Figure 7.13 shows the predicted intelligibility scores using STOI at a range of SNRs for TVGP, and TVDIS based modified speech compared to ORG speech in (a) BMN and (b) SSN, respectively. Each point in the figure denotes scores averaged across utterances for a certain SNR and masker. The STOI scores of TVGP and TVDIS modified speech are both consistently higher than those of the original noise speech for both masker. The STOI values of both TVGP and TVDIS modified speech are almost the same for within each maskers across SNRs. However, the STOI scores in the BMN masker are less than those obtained in the SSN masker for all SNRs. In the SSN masker case, the effect of enhancement increases as the level of SNRs getting higher.

Table 7.2 demonstrates the predicted intelligibility scores using GP % at a range of SNRs for TVGP, and TVDIS based modified speech compared to ORG speech in BMN and SSN, respectively. It is evident that the GP scores was higher for modified speech in comparison to the ORG across SNRs for both maskers. On average, the GP scores in the BMN masker were: 31.5%, 39.2%, and 37.7% for ORG, TVGP and TVDIS, respectively.

In addition, in the SSN masker the average GP score were: 23.8%, 33.7%, and 31.5% for ORG, TVGP and TVDIS, respectively. The average intelligibility estimate increased by around 10% in the SSN for TVGP and TVDIS modified speech compared to the ORG. By contrast the increase was lower in BMN masker with roughly 7.7% and 5.7% for TVGP and TVDIS modified speech compared to the ORG.

The findings seem to support that the modified speech improve the objective intelligibility of speech compared to the ORG. The STOI and GP metric agreed that the TVGP and TVDIS modified speech more intelligible than the ORG. Nevertheless, the STOI measure did not give an clear vision on which modified speech is more intelligible than the ORG, whereas the GP metric predicted that the TVGP modified speech gave more intelligibility estimate than the TVDIS modified speech.

7.5.5 Performance analysis of applying the dynamic range compression on the original speech

In this experiment, we study the effect of the time-domain dynamic range compression on the original data using STOI and GP.

The dynamic range compression (DRC) method was applied as in Zorila et al. (2012) (see Figure 7.7). The method was motivated by the compression techniques in audio broadcasting and hearing-aid amplification Blesser (1969). It works by reallocating energy over time domain in order to reduce the signal's temporal envelope. Thereby, low-energy parts of speech signals (*e.g.,* nasals, onsets and offsets) are amplified, while more energetic voiced sounds are attenuated. In particular, the speech samples are re-scaled using time-varying gains $g(t)$ that are calculated from the dynamically and statically compressed temporal envelope using the following steps according to Zorila et al. (2012):

1. we estimated the temporal envelope by calculating the magnitude of analytical signal from speech signal. The resulting estimated envelope signal was divided into non-overlapping segments of 2.5 $\times$ talker's mean pitch period, then 95% of the maximum value in each individual frame was saved to produce $e(t)$.

2. we dynamically compressed the $e(t)$ with 2 ms release and an instantaneous attack time constants. In the static stage, 30% of the maximum value of dynamically compressed envelope $\hat{e}(t)$ was used as reference level to convert this signal in dB

3. then we apply a predefined input/output envelope characteristic (IOEC)

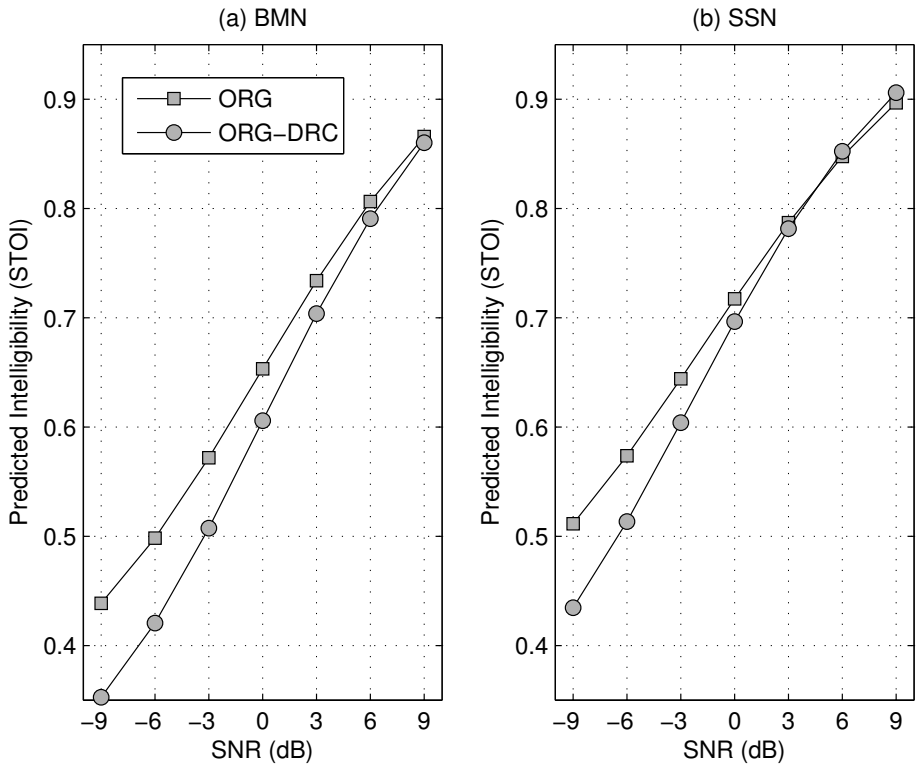

Figure 7.14: The predicted intelligibility using STOI as a measure of speech intelligibility to study the impact of DRC on ORG speech in: (a) BMN and (b) SSN at a range of SNRs.

4. we used the gains obtained from the previous operations in order to rescale the speech samples.

5. finally we applied a global power correction in order to keep the same energy of the speech signal before and after implementing DRC.

Figure 7.14 shows the predicted intelligibility using STOI as a measure of speech intelligibility to study the impact of DRC on ORG speech in: (a) BMN and (b) SSN at a range of SNRs. Here, we chose the reference signal was the original (ORG) and the distorted signal was either the noisy ORG or the noisy ORG-DRC (*i.e.,* the original speech processed with the DRC). According to STOI, the ORG-DRC was more distorted than the ORG speech and precisely at lower SNRs level (*i.e.,* −9, −6, and −3 dB). We Further investigated the impact of the DRC on the ORG using the GP metric as an average across the same range of SNRs used in the STOI within each masker separately as shown in Figure 7.15. It can be seen that the average intelligibility increased by

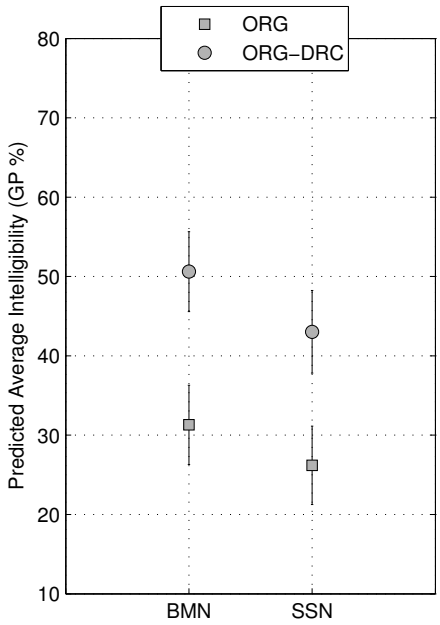

Figure 7.15: The predicted intelligibility using average GP as a measure of speech intelligibility to study the impact of DRC on ORG speech in BMN and SSN maskers.

approximately 20% in the both masker case for ORG-DRC compared to ORG.

The findings here appear contradictory in that they the predict which speech type was more intelligible. In particular, the STOI measure indicated that the ORG-DRC less intelligible than the ORG, whereas the GP predicted that the ORG-DRC is more intelligible than the ORG. This is might be better explained by the fact that the DRC method increase the number of glimpses but the method introduced more distortion to the signal.

7.5.6 Performance analysis of applying the dynamic range compression on the modified speech

In the previous section, we investigated the effect of DRC on the ORG speech type using the objective measure. In this experiment, we further examine the impact on the modified speech including TVGP and TVDIS using the same criteria for evaluation.

Figure 7.16 reveals that the predicted intelligibility using the average score of (a) STOI and (b) GP to study the impact of DRC on modified speech (TVGP and TVDIS) in BMN and SSN maskers, respectively. In order to calculate the STOI, we chose the reference signal to be the modified signal processed by either TVGP or TVDIS and the

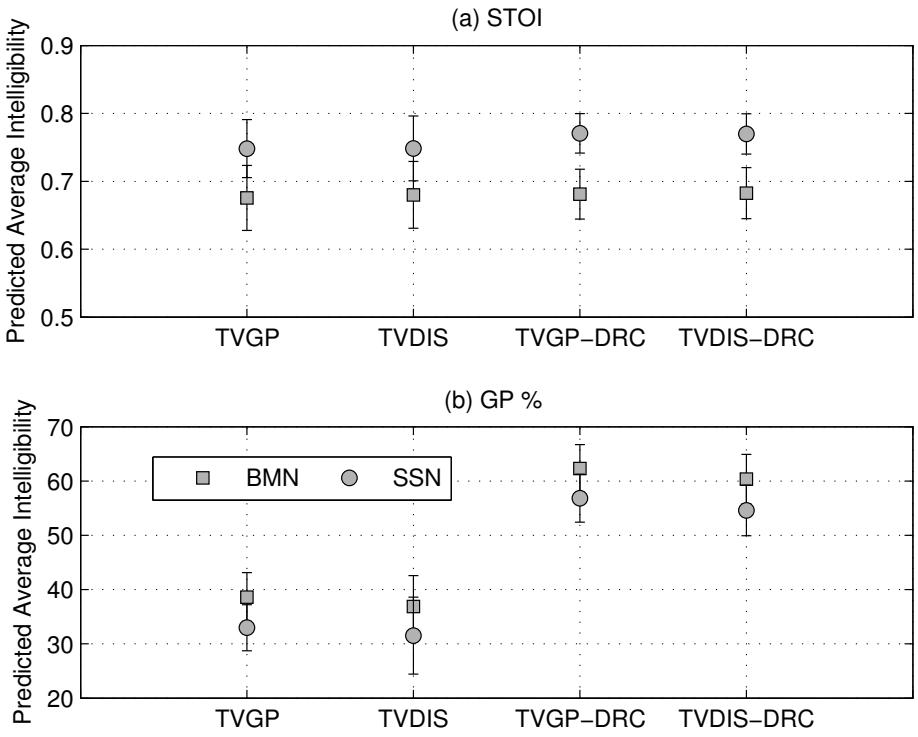

Figure 7.16: The predicted intelligibility using the average score of (a) STOI and (b) GP to study the impact of DRC on modified speech (TVGP and TVDIS) in BMN and SSN maskers.

distorted signal is the noisy modified speech after applying the DRC (*i.e.,* TVGP-DRC and TVDIS-DRC). In line with the finding obtained from the previous experiment, the average GP scores seemed to predict that the applying DRC on the modified speech results in more intelligibility estimate than without DRC. What is surprising is that the STOI results reported the same or a slight increase when implemented the DRC on the modified speech. This may indicate that time-domaine modification is more likely to affect the operation of the STOI and thus it failed to give an accurate indication of the intelligibility estimate.

An example is clearly illustrated in Figure 7.17. The figure shows the clean (Spectro-Temporal Excitation Pattern) STEP representations of both TVGP and TVDIS modified speech with and without dynamic range compression (DRC) compared to the ORG speech of a male talker speaking 'bin green at k z now'. Note that both TVGP and TVDIS modified speech were optimised for BMN masker. The STEP was produced by filtering the speech signal using a bank of 32 gammatone filters with centre frequencies

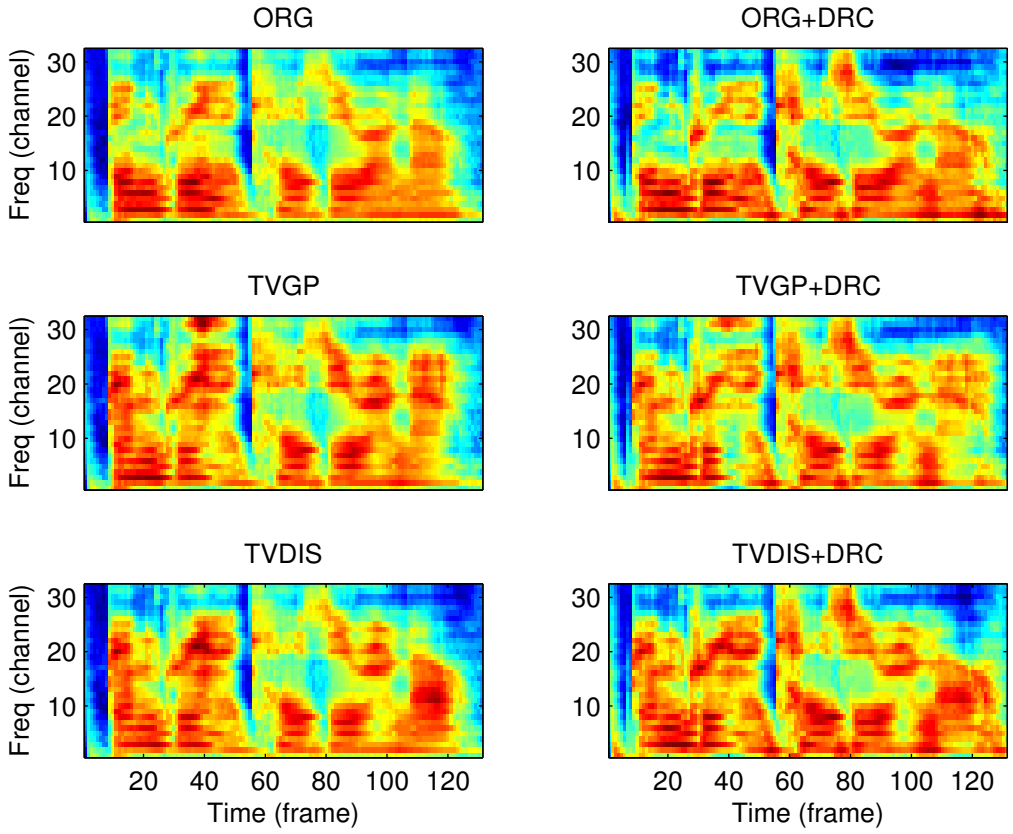

Figure 7.17: An illustration of the clean STEP representations of both TVGP and TVDIS modified speech with and without dynamic range compression (DRC) compared to the ORG speech of a male talker speaking 'bin green at k z now'. Note that both TVGP and TVDIS modified speech were optimised for *babble-modulated* masker. The STEP was produced by filtering the speech signal using a bank of 32 gammatone filters with centre frequencies spaced on an ERB-rate scale from 50 to 8000 Hz, then sampling the low-pass filtered Hilbert envelope at the output of each filter to 100 Hz and log-compressed.

spaced on an ERB-rate scale from 50 to 8000 Hz, then sampling the low-pass filtered Hilbert envelope at the output of each filter to 100 Hz and log-compressed. It is apparent that low-energy segments of speech (*e.g.,* nasals, onsets and offsets) are amplified, while more energetic voiced sounds are attenuated when applied the DRC method on the ORG, TVGP and TVDIS speech types, respectively.

7.6 Human Listening Experiment

In the previous section, we investigate the performance of the modification systems in steady-state and fluctuating masker, and when combined the dynamic range compression using either average relative spectra or objective intelligibility measures.

The aim in this section is to validate the effects of the near-end enhancement methods, developed in this chapter, on the intelligibility of speech in the presence of two type of maskers compared to original unmodified speech (ORG) using human listeners. The methods under investigation includes (see Table 7.1): (i) TVGP-DRC, (ii) TVDIS-DRC, and finally, as reference (iii) SSDRC represented the state-of-the-art method. The maskers were the steady-state speech-shaped noise (SSN) and the babble modulated noise (BMN). For details about the speech and noise materials alongside with how to generate the stimuli see Section 7.5.1.

7.6.1 Participants

Twenty four normal-hearing subjects participated in the study. Listeners were students and staff at the University of Sheffield whose age ranged from 18 to 30 years. The listeners were required to be native English speakers, with no history of speech and/or language dis-orders. All were paid for their participation. Ethics permission was obtained following the University of Sheffield Ethics Procedure.

7.6.2 Procedure

The four speech types namely: ORG, TVGP-DRC, TVDIS-DRC, and SSDRC, were tested in 3 SNRs conditions of the 2 maskers using a total of $19,584$ stimuli (4 speech types $\times$ 816 utterances (34 speakers $\times$ 24 utterances) $\times$ 6 noise conditions (2 maskers $\times$ 3 SNRs)) divided into independent blocks of 136. The independent block was drawn at random, without replacement in which a single subject would hear 34 utterances from each speech types into 6 blocks. The subjects were assigned into blocks in which:

1. each subject heard one block of 136 (34 utterances $\times$ 4 speech types) utterances in each of the 6 noise conditions;

2. no subject heard the same utterance twice;

3. each noise condition was heard by the same number of subject.

The twenty four subjects were divided into two groups of 12. The former group of participants were initially presented with a 3 blocks in the SSN and the remaining 3 blocks in the BMN. The latter group of participants were presented with a reverse order of blocks in which the subject heard a 3 blocks in the BMN first and then a 3 blocks in the SSN.

Subjects were tested individually in an IAC single-walled acoustically-isolated booth. Stimuli were presented once only, and subjects were not able to change the previous output. Noisy utterances were scaled to generate a presentation level of approximately 68 dB SPL. The task was to identify the **letter** and **digit** spoken and type the heard keywords. Once a participant had typed a response, the subsequent stimulus was presented automatically. Null responses were not permitted. Stimulus presentation and response collection was under a web-based experiment system. A subject test took on average in 50-60 minutes. To familiarise them with the test procedure, subjects were given a training session at the beginning of the experiment by listening to and reporting on clean sentences. Subjects were unable to modify the output level.

7.6.3 Results and discussion

In general, performance across listeners was reasonably consistent, so only the mean of the actual identification rates with standard errors averaged across listeners as a function of SNR in the two maskers and the four speech types are plotted in Figure 7.18. It is evident that the intelligibility of modified speech was substantially higher than that achieved by human listeners listening to ORG corrupted speech across the SNRs for both maskers.

A two-way repeated measure ANOVA with two within-subjects factors (modification type and masker type) on arcsine-transformed identification rates indicated statistically significant main effects of modification type ($F(3,15) = 21.5$, $p < 0.001$), and the SNR level across maskers ($F(5,15) = 22.7$, $p < 0.001$) on the actual intelligibility of speech in noise. This suggests that the effect of modification strategies varied across SNR and were significantly different from ORG for each masker type.

A post hoc test according to Fisher's LSD ($\alpha = 0.05$), computed separately for each masker type across the SNR level using ANOVAs with the single factor of modification type, indicated several significant differences between the different experimental conditions. The p-values can be found in Table 7.3.

Mean intelligibilities were 46.5 %, 70.5 %, 60.3 % and 54.4 % for ORG, TVDIS-DRC, SSDRC and TVGP-DRC across SNRs levels for BMN masker, respectively. Furthermore,

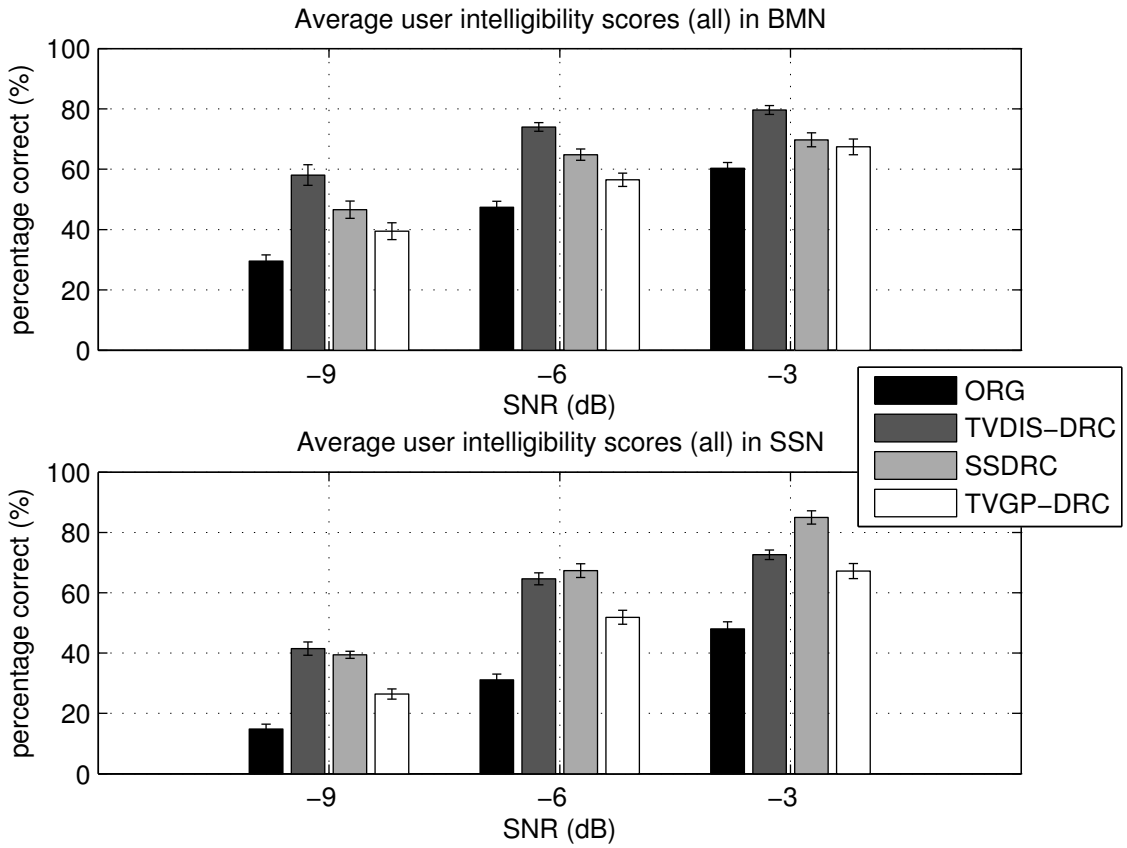

Figure 7.18: Percentage of correct identifications of both letter and digit in each speech type with the standard errors in the babble-modulated (upper panel) and speech-shaped noise (lower panel). The *black* shows performance in the ORG entry as a function of the signal-to-noise-ratio (SNR) used to produce the stimulus. The *dark grey* in the figure shows performance in the TVDIS-DRC entry as a function of the SNR used to produce the stimulus. The *light grey* in the figure shows performance in the SSDRC entry as a function of the SNR used to produce the stimulus. The *white* shows performance in the TVGP-DRC entry as a function of the SNR used to produce the stimulus. Note that all were presented at the same overall RMS level.

mean intelligibilities were 31.3 %, 60.5 %, 64.3 % and 48.5 % for ORG, TVDIS-DRC, SSDRC and TVGP-DRC across SNRs levels for SSN masker, respectively. The most striking result to emerge from the data is that, in all speech type, the amount of increase on average intelligibility in the BMN masker was more than that in the SSN masker.

The intelligibility gain varied across modification types, maskers, and SNR levels. In general, the intelligibility gains were slightly higher for modified speech in the presence of the SSN masker compared to the ORG than in the BMN masker. The TVDIS-DRC

Table 7.3: p-values for comparing intelligibility scores between systems across maskers.

Masker	Methods	ORG	TVDIS-DRC	SSDRC	TVGP-DRC
BMN	ORG	-	0.111	0.145	0.020
	TVDIS-DRC	0.111	-	0.033	0.091
	SSDRC	0.145	0.033	-	0.124
	TVGP-DRC	0.020	0.091	0.124	-
SSN	ORG	-	0.182	0.089	0.098
	TVDIS-DRC	0.182	-	0.092	0.084
	SSDRC	0.089	0.092	-	0.009
	TVGP-DRC	0.098	0.084	0.009	-

method outperformed all other speech types across SNRs level in BMN. Additionally, the intelligibility gain were roughly the same for the TVGP-DRC and SSDRC. In the SSN masker condition, however, the TVDIS-DRC and SSDRC had a similar pattern of increase at -9 and -6 dB SNR, but the SSDRC (85 %) outperformed the TVDIS-DRC (72 %) at higher level of SNR. The lowest intelligibility gains were obtained by TVGP-DRC in both maskers and was noticeable in the SSN masker at all level of SNRs.

All speech types were preprocessed by different spectral modification methods and then by the same time-domain modification namely DRC method. In the large scale evaluation of modification methods in noise that carried out in 2013 by Cooke et al. (2013b), they found that the modification method that combined the time-domain modification namely DRC, named SSDRC in this evaluation, significantly outperformed all other modification methods across both steady-state and completing-speaker masker for most SNRs level specifically at the lower SNRs. In our evaluation, we inferred that combing the time-domain modification with the time-varying spectral modification resulted in higher intelligibility gain. We also akin the higher gain of intelligibility obtained by TVDIS-DRC to using a better intelligibility-optimisation method that optimised a phoneme-level discriminative microscopic intelligibility. These findings suggested that a significant gain can be achieved by first defining better objective intelligibility measure, and second by combing time-domain modification method.

We further tested the actual identification rates (%) against the predictive score obtained by objective measures including GP (%) and STOI. Figure 7.19 shows the results of Scatter plots of GP predictions versus actual intelligibility scores and STOI predictions versus actual intelligibility scores in BMN and SSN masker. Each point denotes scores averaged across utterances and listeners for BMN and SSN maskers separately.

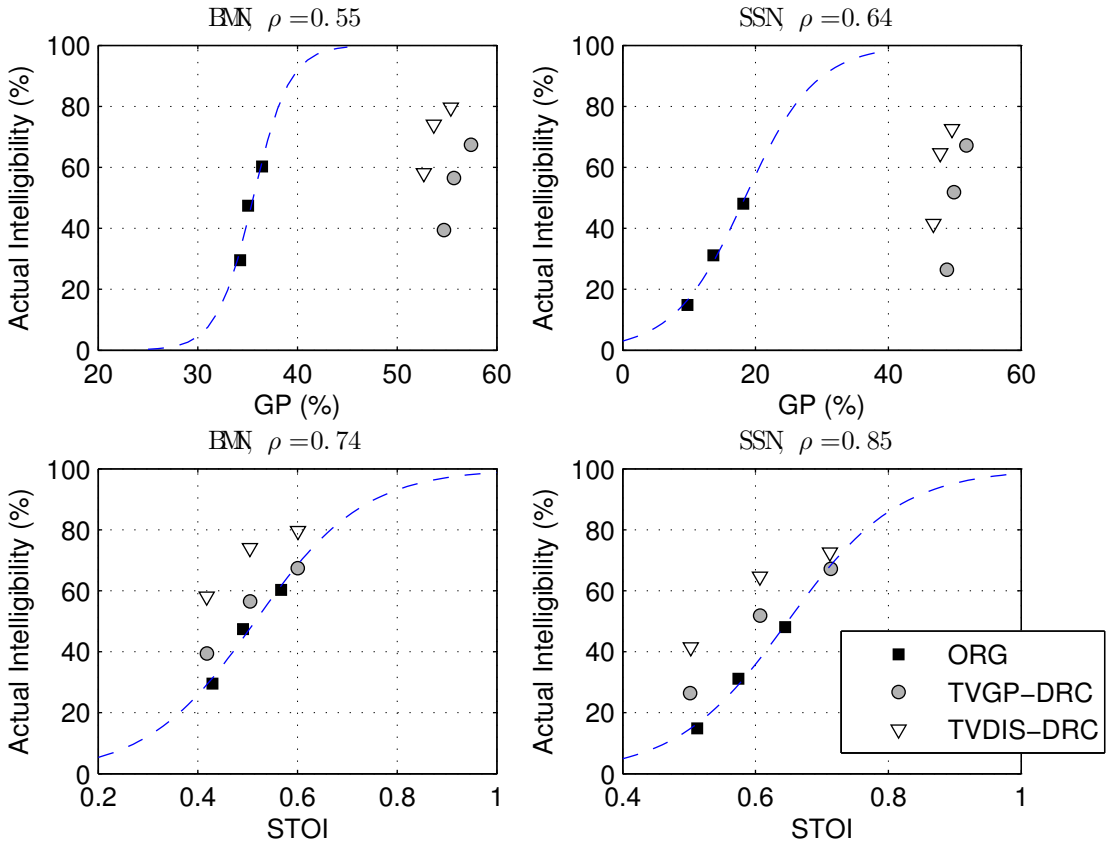

Figure 7.19: Scatter plots of - at *left* panels - GP predictions versus actual intelligibility scores and STOI predictions versus actual intelligibility scores in BMN masker, and - at *right* panels - GP predictions versus actual intelligibility scores and STOI predictions versus actual intelligibility scores in SSN masker.

It can be seen that the STOI measure had a high level of agreement with the actual listening results compared to the GP measure for both maskers $\rho = 0.74$, and $\rho = 0.85$, respectively. Furthermore, a lower level of correlation were obtained in the GP case for both maskers compared to the STOI case as shown in the figure. The plot also shows the best fitting logistic function to map the predictions scores to actual intelligibility scores.

7.7 General Discussion

In this thesis, there has been considerable interest in the potential relevance of the proposed discriminative microscopic intelligibility model 'DIS' to improve speech pre-

enhancement systems and speech intelligibility. The system proposed in the previous chapter, in which the DIS model was optimised, failed to fulfil this objective. It was hypothesised that the reason for getting a below chance performance is that a time-invariant spectral shaping method lacks the precision needed for handling rapidly time-varying speech signals. A time-varying spectral shaping method would be more appropriate to apply within the DIS model. Thus, a time-varying spectral shaping method has been proposed here and applied on a segment-by-segment basis, aiming to account for the masker's temporal fluctuations. The method defines a spectral weighting pattern for each S-T element in order to emphasise those important S-T elements above the level of the masker at a segmental level, enabling temporal variations between segments to be tracked by the listener. We hypothesise that applying a temporally-varying spectral shaping method to the target speech is likely to be most beneficial when the masker itself is also modulated, since it helps to define an accurate weight for each S-T element.

The findings showed that the proposed systems, *i.e.,* TVGP and TVDIS, have led to substantial redistribution of energy in the average relative spectra from lower to middle and higher frequencies for sentences presented in BMN masker relative to an unmodified ORG speech (see Figure 7.8). The TVDIS was further compared to DIS-OPT in SSN masker (as illustrated in Figure 7.10). The results suggested that there was an obvious benefit from applying the TVDIS system in comparison to the DIS-OPT where no evident benefit could be observed. We also found a significant gender-related difference in the average relative spectra and the DC factors such that the male talkers in the Grid corpus would receive an advantage from the TVDIS system more than the female talkers (as shown in Figure 7.12). In the human listening test, we found significant increases in intelligibility for sentences presented in both BMN and SSN masker when combining the time-domain modification, *i.e.,* DRC, with the TVGP and TVDIS systems compared to ORG speech baseline. The most successful technique evaluated subjectively in BMN masker was TVDIS-DRC, producing increases in identification scores over the TVGP-DRC and SSDRC. The DRC component of TVDIS-DRC, proposed in Zorila et al. (2012), works by transferring energy from sonorant to less sonorant parts of speech signal. This results in the enhancement of the transient components of speech which may explain its ability to improve intelligibility of modified speech in noisy conditions.

Using a speech model from missing data ASR has an implication on the complexity of the near-end intelligibility enhancement system. The missing data ASR uses speaker-dependent phoneme models. As a result, the number of models grows from 39 (the number of phonemes in the English language) to 39 $\times$ 34 (the total number of speakers

in the Grid corpus). This represents 39 phonemes models as unique versions to each speaker. To reduce such complexity, the phoneme models are represented as single-state (rather that the usual 3-states used in ASR systems), and thus the computation of the DIS model is simplified. It is calculated based on the acoustic features within each time frame at a defined segment using the the single-state SD phonemes models.

A second implication of the proposed approach is the segmentation and its possible audible artefacts that may occur in the resynthesised modified speech. Care needs to be taken to avoid tonal artefacts and discontinuities at the boundaries between segments. Therefore, the time-varying spectral modification is applied with a linear interpolation ranging from no-shaping to maximum shaping at the central time-frames and back to no-shaping for a pre-defined segment. To improve the optimality of the pre-enhancement system, the multiple segments should be optimised simultaneously. Petkov et al. (2013) applied a similar procedure where the segmentation boundaries were based on the word boundaries and the optimisation was implemented individually for each word. However, there are often no pauses between words. So, in our system, to minimise the effects of discontinuity at segment boundaries we choose to place boundaries at the low energy points in the speech signal.

One limitation of the system is the lack of consistency in the findings across the SNRs which might be associated to the optimisation algorithm. Although the the Nelder-Mead algorithm (see Section 4.4.3 in Chapter 4) is appropriate for finding a better solution for the unconstrained problem, it estimates a local maxima based on the current estimates of the simplex. The size and position of the simplex is changing within each alteration of the optimisation which might not alway guarantee the optimal local maxima. An additional shortcoming is the listening evaluation process using headphone presentation of speech and noise in an acoustically-isolated booth. The listening evaluation process does not reflect the situation in which listeners will have access to speech in a real environments. This is a very challenging task because of other sound sources and cues available in that environments.

7.8 Chapter Summary

To conclude, we presented a time-varying spectral modification approach for improving the intelligibility in noise. The speech spectrum was first divided into spectro-temporal segments. A spectral shaping function was generated for each time-frame of each segment by cepstral coefficients. This was achieved by optimising a measure of energetic

masking - as a baseline system - and a discriminative microscopic intelligibility measure defined in the previous chapter. The technique was objectively and subjectively evaluated using in both stationary and non-stationary masking conditions and over a wide range of SNRS. By combining dynamic range compression with the proposed method, we managed to achieve high intelligibility scores than those achieved by the well known SS-DRC approach.

Chapter 8

Conclusions and Future Work

This chapter provides a summary of this thesis in Section 8.1, main contributions in Section 8.2, and finally presents some suggestions for future work in Section 8.3.

8.1 Summary of the Thesis

We set out to develop a closed-loop optimisation approach to near-end intelligibility enhancement which works by exploiting a *priori* knowledge of a speaker and the noise environment to increase the intelligibility of speech in noise. We started by building a general framework for the closed-loop feedback system using the analysis-resynthesis system described by Hohmann (2002), an optimisation procedure, and a measure of speech intelligibility. We then automatically modified the speech signal according to the environmental noise by maximising the intelligibility estimate without changing the energy level of speech. The main advantage of this closed-loop approach is the availability of a feedback signal. To achieve the aim of exploiting a priori knowledge of the speaker, we derived a novel discriminative intelligibility measure based on a statistical model of speech from the speaker that was to be enhanced. Specifically, we employed a speaker-specific hidden Markov model and considered the ratio of the likelihoods of the correct state sequence and the best scoring competing state sequence using missing feature theory to account for masking. We observed in listening tests that a simple stationary spectral modification based on maximising a measure of energetic masking, the Glimpse Proportion measure (Cooke, 2006), significantly increased intelligibility in stationary noise conditions. However, coupling the discriminative intelligibility model with the simple stationary spectral modification method did not improve intelligibility in the same stationary noise conditions (see Figure 6.10). We thus enhanced the system

by: (i) extending the discriminative intelligibility model to work at a phoneme-level, (ii) proposing a time-varying spectral shaping, and (iii) performing the optimisation on a segment-by-segment basis. Results showed that combining this system with a time-domain noise independent method (*i.e.,* dynamic range compression) improved intelligibility particularly in non-stationary noise when compared to the state-of-the-art noise-independent system, *i.e.,* spectral shaping and dynamic range compression, (see Figure 7.18).

8.2 Contributions

In this section, we will present the main contributions of this thesis to the different areas of knowledge: speech intelligibility modelling, methods for modification strategies, near-end intelligibility enhancement systems, and perception of modified speech in noise.

8.2.1 Intelligibility modelling

The starting point of this work was the selection of an appropriate intelligibility model for a good design of a closed-loop near-end intelligibility enhancement system. We began with the Glimpse Proportion (GP) measure which was based on an auditory masking model using speech and noise signals. We then defined a new discriminative intelligibility model using missing data theory. Thus the main contribution to the field of intelligibility modelling can be listed:

- **The development of a new discriminative microscopic Intelligibility model.**
  We proposed a discriminative microscopic Intelligibility (DIS) model aiming at minimising the confusions between acoustically similar speech units (see Section 6.2). The DIS measure used extra knowledge compared to the GP measure in which a pre-trained clean speech model is required to treat the missing feature of the mask using the reliable feature. We derived the DIS from a statistical model of speech from the speaker that is to be enhanced. Specifically, we employed a hidden Markov model and considered the ratio of the likelihoods of the correct state sequence and the best scoring competing state sequence using missing feature theory to account for masking. We claimed that the GP measure accounts for the audible spectro-temporal elements while the DIS accounts for the audible and most distinguishable spectro-temporal elements when masking occurred.

However, this model did not show a significant intelligibility improvement when implemented in the close-loop framework for near-end intelligibility enhancement compared to the GP measure. Based on that we decided to extend the definition of this measure, in Chapter 7, to work at a phoneme level. So rather than having an average estimate that represents the probability of correct word sequence against a single candidate (most erroneous) word sequence, we had an average estimate between the probability of the correct phones sequence and most erroneous phones sequence of that utterance. Employing the extended version of this model in the closed-loop based near-end intelligibility enhancement system provided greater intelligibility gains than when using the GP measure in such system (see Figure 7.18). The contribution of this measure differed from the measure proposed by Petkov et al. (2013) in which they optimised the probability of correct recognition of the transmitted message.

8.2.2 Methods for modification strategies

With respect to the second key element of this thesis which was designing flexible speech modification strategies, we have made the following contributions.

- **The development of a spectral modification method.**
  Inspired by the empirically-observed characteristics of natural speech produced in noise (*i.e.,* Lombard speech), we proposed a spectral shaping method in Chapter 5 (Section 5.3). We computed the spectral modification based on modifying the first few coefficients of an auditory cepstral representation to increase intelligibility of modified speech in noise. To reduce the complexity of this method and to limit decreasing the quality of the modified speech, we proposed to use just the first few parameters.

  We found that implementing this method in an optimisation based framework with the GP measure did indeed increase intelligibility of speech in stationary noise.

- **The development of a time-varying spectral shaping method.**
  We extended the idea used in the stationary-spectral shaping method to account for time-varying changes. Thus we proposed in Chapter 7, a time-varying spectral shaping method operated by defining a weight for each spectro-temporal element of the spectrum through modifying the cepstral coefficients ranging from no-shaping (at the beginning and the end frames of the spectrum) to the actual shaping in the middle frames.

To improve the performance of this method, we proposed to implement the time-varying spectral shaping on a segment-by-segment basis. Segment boundaries were selected by locating minima in the time-varying energy of the speech signal.

This method was applied in the closed-loop feedback based near-end intelligibility enhancement system and proved its effectiveness in improving speech intelligibility in stationary and non-stationary noises. The improvement was greater when optimising the phoneme-level DIS compared to when optimising the GP. Larger intelligibility gains were observed when time-domain dynamic range compression was also applied.

8.2.3 Near-end intelligibility enhancement systems

Throughout this thesis, we developed and evaluated several closed-loop based near-end intelligibility enhancement systems to increase the intelligibility of speech in noise. These systems work by choosing an objective intelligibility measure and a modification strategy under an energy preservation constraint. These systems are listed below along with an indication of the chosen objective intelligibility measure, the modification strategy, and their effectiveness in improving intelligibility in noise:

- **GP-based spectral modification system.**
  As a baseline system, we used the GP measure and the stationary spectral modification strategy in Chapter 5. This system operates by manipulating the first few coefficients of an auditory cepstral representation such as to maximise the GP measure at utterance level using speech and noise signals. Results showed this system improved speech intelligibility in a stationary noise masker over a range of SNRs: -9, -6, -3, 0, and 3 dB.

- **DIS-based spectral modification system.**
  Chapter 6 coupled a new defined intelligibility model, DIS, with the stationary spectral modification method. In this system, the DIS was maximised to better tune the cepstral parameters at an utterance level. This system required knowledge about the speech and noise signals and a model of speech priori to processing. We tested this system with a stationary noise masker at different noise levels. Results showed that this system did not lead to significant increases in intelligibility compared to the original unmodified speech and was not as effective as the GP-based spectral modification system.

The development and subjective evaluation of the above systems was published in Al Dabel and Barker (2014) and Al Dabel and Barker (2015), respectively.

- **GP-based time-varying spectral modification system.**
  In Chapter 7, the GP measure was coupled with a new time-varying spectral shaping that operates on a segment-by-segment basis. As mentioned earlier, segment boundaries were chosen based on locating minima in the time-varying energy of the speech signal. Weights were set by optimising the GP while using a priori knowledge of the speech and noise signals. Objective evaluation using the short-term objective intelligibility (STOI) (Taal et al., 2011) demonstrated a higher intelligibility gain. In human listening tests, we combined the output of this system with the time-scale modification method (*i.e.,* the dynamic range compression system) for a fair comparison to the reference system (spectral shaping and dynamic range compression) proposed by Zorila et al. (2012). Results showed significant and consistent improvements in subjective intelligibility for the modified speech compared to the original speech in both stationary and non-stationary noises at various level of noise. However, the intelligibility gain was less than that achieved by the reference system.

- **DIS-based time-varying spectral modification system.**
  In Chapter 7, we grouped the phoneme-level DIS measure with the time-varying spectral shaping on a segment-by-segment basis. This system required speech and noise signals, as well as a speaker-specific model of speech. This system was also combined with the dynamic range compression for the subjective listening test. Results demonstrated that the DIS-based time-varying spectral shaping combined with the dynamic range compression outperformed the reference system in the non-stationary noise but there was not significant improvement compared to the reference system in the stationary noise.

  We conducted an across-speaker analysis to try and relate speaker specific intelligibility gains to the speakers' original intelligibility, *i.e.,* do speakers who are poorly intelligible gain more from the enhancement. We found that the male speakers were more likely to benefit from this system in noise compared to the female speaker.

8.2.4 Perception of modified speech in noise

- **Human listening experiment (1).**
  Our first subjective experiment was presented in Chapter 5 and 6. We assessed the two closed-loop near-end intelligibility enhancement systems developed in these chapters and compared them to original unmodified speech and a reference noise-independent spectral shaping system. The evaluation was carried out in stationary noise. We found that for intelligibility to be improved the GP-based spectral modification system was more effective than the DIS-based spectral modification system, and was comparable to the reference system in terms of intelligibility improvement.

- **Human listening experiment (2).**
  Our final subjective experiment in Chapter 7 was performed to evaluate two closed-loop based near-end intelligibility enhancement systems developed in this chapter combined with time-domain noise-independent method (*i.e.,* dynamic range compression). We also compared these systems with the original unmodified case and a reference noise-independent method in stationary and non-stationary noises. The results showed that combing dynamic range compression with the DIS-based time-varying spectral shaping system increased intelligibility, particularly in non-stationary noise.

8.3 Directions for Future Work

The research on developing new near-end intelligibility enhancement to accommodate various application scenarios is not concluded. Based on our experience of systems developed in this thesis, we envision the following directions for future research:

- **Possible future developments for addressing speaker-related difficulties.**
  The novel DIS-based time-varying spectral modification system has been demonstrated to work well in stationary and non-stationary noises and more precisely for male speakers. The fact that this system is a speaker-dependent system may offer us the opportunity to extend it to speakers with special needs (*i.e.,* speakers with speech disorders). A possible approach could be to personalise the phoneme-level DIS model by using a statistical model of a speaker to whom the enhancement is designed for. Although the possible applications that may embed such enhancement

could be limited, it is still potentially useful particularly in medical applications and assistive technology.

The DIS measure by its own could also be extended to be a measure of confusability for a pathological speech intelligibility assessment. It may help clinicians to identify speech units which can be expected to be easily confused with other speech units.

- **Possible future developments for addressing listener-related difficulties.** The GP-based and DIS-based near-end intelligibility enhancement systems developed in this thesis cloud be advanced to accommodate people with hearing impairment or cochlear hearing loss. Since people with hearing-loss often have auditory filters that are broader than those in people with normal-hearing (Glasberg and Moore, 1986), one could extend the modification strategy to model frequency selectivity.

  An improved understanding of the relevant mechanisms of hearing loss might suggest further advances in near-end intelligibility enhancement techniques, in speech intelligibility prediction models, or in modification strategies for assistive listening devices such as hearing aids and cochlear implants tuned to be listener-dependent or to accommodate specific group of listeners.

- **Possible future developments for addressing environment-related difficulties.**
  To apply the closed-loop near-end intelligibility enhancement systems in reverberation, one could model the effect of late reverberation on speech and thus developed time-varying spectral shaping to minimise such effect. We anticipate that the GP-based and DIS-based systems may work well in the reverberant environment by minimising the overlap-masking, which often occurs when the energy of a phoneme masks the phonemes that follow. Also, the DIS model may benefit from overlap-masking effect that may cause phonetic confusions, the DIS thus is more likely to improve the discriminability.

- **Possible future developments for relaxing system constraints.**
  Possible extensions that can follow the work from this thesis may include further analysis of the following:

  i) the quality of enhanced modified speech using quality related listening tests,

ii) applying the closed-loop systems under a loudness constraint using a loudness model of Moore and Glasberg (1996) or Moore et al. (1997) (example work is in Valentini-Botinhao et al. (2013)),

iii) applying the closed-loop systems using other possible modification strategies,

iv) applying different segmentation strategies (*e.g.,* vowel centres, words boundaries),

v) applying time-varying strategies to the spectral modification

vi) using different statistical modelling techniques to represent the speech signal (*e.g.,* deep neural network (DNN)).

- **Possible future developments for the complexity of the system.**
  Reducing the complexity of the optimisation-based near-end intelligibility enhancement systems is a trad-off between optimising the parameters and the benefits they gave. Thus, more experiments are required in order to better understand the trade-off better and hence trying to reduce the number of parameters without any further cost in the performance. One could also generalise the closed-loop near-end intelligibility enhancement system to different datasets.

- **Possible future developments for speech modification.**
  The modification strategies developed in this thesis are applied to shape speech signal. There is an opportunity to define more sophisticated speech modifications to work in an enhancement-by-resynthesis framework (*e.g.,* Carmona et al. (2013)). The enhancement-by-resynthesis framework often use the HMM to decode speech into a model and state sequence which is then input into an HMM-based speech synthesis to output a clean speech signal. In this framework, one could develop several modification strategies in HMM space such timing (*e.g.,* making a vowel shorter, or a phoneme longer), or reorganising and changing the actual words sequence.

With this thesis, it is our hope that we have contributed knowledge to new approaches for enhancing speech intelligibility in noise, and that we have provided a useful closed-loop framework that inspire others to consider this approach to near-end intelligibility enhancement. In addition, we hope that we have contributed a fresh look at using missing data theory in intelligibility modelling and near-end intelligibility enhancement.

References

Ainsworth, W. and Meyer, G. (1994). Recognition of plosive syllables in noise: Comparison of an auditory model with human performance. *The Journal of the Acoustical Society of America*, 96(2):687–694.

Al Dabel, M. and Barker, J. (2014). Speech pre-enhancement using a discriminative microscopic intelligibility model. In *Proc. Interspeech*, pages 2068–2072, Singapore, Singapore.

Al Dabel, M. and Barker, J. (2015). On the role of discriminative intelligibility model for speech intelligibility enhancement. In *Proc. ICPHS XVIII*, Glasgow, UK.

Allen, J. (2005a). *Articulation and intelligibility*. Morgan & Claypool, San Rafael, CA.

Allen, J. (2005b). Consonant recognition and the articulation index. *The Journal of the Acoustical Society of America*, 117(4):2212–2223.

Amano-Kusumoto, A. and Hosom, J. (2010). A review of research on speech intelligibility and correlations with acoustic features. Technical report CSLU-11-002, Department of Biomedical Engineering, Oregon Health & Science University.

ANSI, A. (1997). S3. 5-1997, methods for the calculation of the speech intelligibility index. *New York: American National Standards Institute*, 19:90–119.

Arai, T., Hodoshima, N., and Yasu, K. (2010). Using steady-state suppression to improve speech intelligibility in reverberant environments for elderly listeners. *IEEE Tran. Audio, Speech, and Language Processing*, 18(7):1775–1780.

Arai, T., Kinoshita, K., Hodoshima, N., Kusumoto, A., and Kitamura, T. (2002). Effects of suppressing steady-state portions of speech on intelligibility in reverberant environments. *Acoustical Science and Technology*, 23(4):229–232.

Arai, T., Yasu, K., and Hodoshima, N. (2004). Effective speech processing for various impaired listeners. In *Proc. the 18th International Congress on Acoustics*, pages 1389–1392.

Assmann, P. and Summerfield, Q. (2004). The perception of speech under adverse conditions. In *Speech processing in the auditory system*, pages 231–308. Springer.

Atal, B. and Hanauer, S. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50(2B):637–655.

Aubanel, V. and Cooke, M. (2013). Information-preserving temporal reallocation of speech in the presence of fluctuating maskers. In *Proc. Interspeech*, pages 3592–3596, Lyon, France.

Aubanel, V., Cooke, M., Villegas, J., and García Lecumberri, M. (2011). Conversing in the presence of a competing conversation: Effects on speech production. In *Proc. Interspeech*, pages 2833–2836, Florence, Italy.

Avendano, C. and Hermansky, H. (1996). Study on the dereverberation of speech based on temporal envelope filtering. In *Proc. ICSLP*, pages 889–892.

Barker, J. and Cooke, M. (2007). Modelling speaker intelligibility in noise. *Speech Communication*, 49(5):402–417.

Barker, J., Cooke, M., and Ellis, D. (2005). Decoding speech in the presence of other sources. *Speech communication*, 45(1):5–25.

Barker, J., Vincent, E., Ma, N., Christensen, H., and Green, P. (2013). The PASCAL CHiME speech separation and recognition challenge. *Computer Speech & Language*, 27(3):621–633.

Blesser, B. (1969). Audio dynamic range compression for minimum perceived distortion. *IEEE Tran. Audio and Electroacoustics*, 17(1):22–32.

Boldt, J. and Ellis, D. (2009). A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation. In *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pages 1849–1853.

Bolt, R. and MacDonald, A. (1949). Theory of speech masking by reverberation. *The Journal of the Acoustical Society of America*, 21(6):577–580.

Bond, Z. and Moore, T. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication*, 14(4):325–337.

Bradley, J., Sato, H., and Picard, M. (2003). On the importance of early reflections for speech in rooms. *The Journal of the Acoustical Society of America*, 113(6):3233–3244.

Bradlow, A. and Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America*, 112(1):272–284.

Bradlow, A., Kraus, N., and Hayes, E. (2003). Speaking clearly for children with learning disabilities: Sentence perception in noise. *Journal of Speech, Language, and Hearing Research*, 46(1):80–97.

Bradlow, A., Torretta, G., and Pisoni, D. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech communication*, 20(3):255–272.

Bregman, A. (1990). *Auditory Scene Analysis: The perceptual organization of sound*. MIT Press.

Brouckxon, H., Verhelst, W., and De Schuymer, B. (2008). Time and frequency dependent amplification for speech intelligibility enhancement in noisy environments. In *Proc. Interspeech*, pages 557–560.

Brungart, D. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 109(3):1101–1109.

Carmona, J., Barker, J., Gomez, A., and Ma, N. (2013). Speech spectral envelope enhancement by HMM-based analysis/resynthesis. *IEEE Signal Processing Lett.*, 20(6):563–566.

Carter, G., Knapp, C., and Nuttall, A. (1973). Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing. *IEEE Tran. Audio and Electroacoustics*, 21(4):337–344.

Chabot-Leclerc, A., Jørgensen, S., and Dau, T. (2014). The role of auditory spectrotemporal modulation filtering and the decision metric for speech intelligibility prediction. *The Journal of the Acoustical Society of America*, 135(6):3502–3512.

Chanda, P. and Park, S. (2007). Speech intelligibility enhancement using tunable equalization filter. In *Proc. ICASSP*, pages 613–616.

Chen, F. and Loizou, P. (2011). Predicting the intelligibility of vocoded speech. *Ear and hearing*, 32(3):331–338.

Christiansen, C., Pedersen, M., and Dau, T. (2010). Prediction of speech intelligibility based on an auditory preprocessing model. *Speech Communication*, 52(7):678–692.

Cooke, M. (1991). *Modelling auditory processing and organisation*. PhD thesis, University of Sheffield.

Cooke, M. (2003). Glimpsing speech. *Journal of Phonetics*, 31(3):579–584.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, 119(3):1562–1573.

Cooke, M. (2009). Discovering consistent word confusions in noise. In *Proc. Interspeech*, pages 1887–1890, Brighton, UK.

Cooke, M., Barker, J., Cunningham, S., and Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424.

Cooke, M., García Lecumberri, M., and Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America*, 123(1):414–427.

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech communication*, 34(3):267–285.

Cooke, M., Hershey, J., and Rennie, S. (2010). Monaural speech separation and recognition challenge. *Computer Speech & Language*, 24(1):1–15.

Cooke, M., King, S., Garnier, M., and Aubanel, V. (2014). The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech & Language*, 28(2):543–571.

Cooke, M. and Lu, Y. (2010). Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. *The Journal of the Acoustical Society of America*, 128(4):2059–2069.

Cooke, M., Mayo, C., and Valentini-Botinhao, C. (2013a). Intelligibility enhancing speech modifications: the Hurricane Challenge. In *Proc. Interspeech*, pages 3552–3556, Lyon, France.

Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., and Tang, Y. (2013b). Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, 55(4):572–585.

Cooke, M., Morris, A., and Green, P. (1997). Missing data techniques for robust speech recognition. In *Proc. ICASSP*, pages 863–866.

Crespo, J. and Hendriks, R. (2014). Speech reinforcement in noisy reverberant environments using a perceptual distortion measure. In *Proc. ICASSP*, pages 910–914.

Darwin, C. and Hukin, R. (2000). Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *The Journal of the Acoustical Society of America*, 107(2):970–977.

Dau, T., Püschel, D., and Kohlrausch, A. (1996). A quantitative model of the "effective" signal processing in the auditory system. I. model structure. *The Journal of the Acoustical Society of America*, 99(6):3615–3622.

Deng, L. and O'Shaughnessy, D. (2003). *Speech processing: a dynamic and optimization-oriented approach*. CRC Press.

Dolson, M. (1986). The phase vocoder: A tutorial. *Computer Music Journal*, 10(4):14–27.

Dreher, J. and O'Neill, J. (1957). Effects of ambient noise on speaker intelligibility for words and phrases. *The Journal of the Acoustical Society of America*, 29(12):1320–1323.

Drugman, T. and Dutoit, T. (2010). Glottal-based analysis of the lombard effect. In *Proc. Interspeech*, pages 2610–2613.

Drullman, R., Festen, J., and Plomp, R. (1994a). Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America*, 95(5):2670–2680.

Drullman, R., Festen, J., and Plomp, R. (1994b). Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America*, 95(2):1053–1064.

Dubbelboer, F. and Houtgast, T. (2007). A detailed study on the effects of noise on speech intelligibility. *The Journal of the Acoustical Society of America*, 122(5):2865–2871.

Dubbelboer, F. and Houtgast, T. (2008). The concept of signal-to-noise ratio in the modulation domain and speech intelligibility. *The Journal of the Acoustical Society of America*, 124(6):3937–3946.

Dudley, H. (1939). Remaking speech. *The Journal of the Acoustical Society of America*, 11(2):169–177.

Durlach, N. (2006). Auditory masking: Need for improved conceptual structurea). *The Journal of the Acoustical Society of America*, 120(4):1787–1790.

Durlach, N., Mason, C., Kidd Jr, G., Arbogast, T., Colburn, H., and Shinn-Cunningham, B. (2003). Note on informational masking (L). *The Journal of the Acoustical Society of America*, 113(6):2984–2987.

Elhilali, M., Chi, T., and Shamma, S. (2003). A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. *Speech communication*, 41(2):331–348.

Ephraim, Y. and Cohen, I. (2005). Recent advancements in speech enhancement. In *The Electrical Engineering Handbook*, pages 15–12, Boca Raton. CRC Press.

Ewert, S. and Dau, T. (2000). Characterizing frequency selectivity for envelope fluctuations. *The Journal of the Acoustical Society of America*, 108(3):1181–1196.

Ferguson, S. and Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 112(1):259–271.

Ferguson, S. and Kewley-Port, D. (2007). Talker differences in clear and conversational speech: acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research*, 50(5):1241–1255.

Festen, J. and Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, 88(4):1725–1736.

Fitzpatrick, M., Kim, J., and Davis, C. (2011). The effect of seeing the interlocutor on speech production in different noise types. In *Proc. Interspeech*, pages 2828–2832, Florence, Italy.

Flanagan, J. and Golden, R. (1966). Phase vocoder. *Bell System Technical Journal*, 45(9):1493–1509.

French, N. and Steinberg, J. (1947). Factors governing the intelligibility of speech sounds. *The Journal of the Acoustical Society of America*, 19(1):90–119.

Freyman, R., Balakrishnan, U., and Helfer, K. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America*, 115(5):2246–2256.

Furui, S. (1986). On the role of spectral transition for speech perception. *The Journal of the Acoustical Society of America*, 80(4):1016–1025.

Gagne, J., Masterson, V., Munhall, K., Bilida, N., and Querengesser, C. (1994). Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech. *Journal-Academy of Rehabilitative Audiology*, 27:135–158.

Gagné, J., Rochette, A., and Charest, M. (2002). Auditory, visual and audiovisual clear speech. *Speech communication*, 37(3):213–230.

García Lecumberri, M. and Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America*, 119(4):2445–2454.

Garnier, M., Bailly, L., Dohen, M., Welby, P., and Lœvenbruck, H. (2006). An acoustic and articulatory study of lombard speech: Global effects on the utterance. In *Proc. ICSLP*, pages 2246–2249, Pittsburgh, USA.

Ghitza, O. (1993). Adequacy of auditory models to predict human internal representation of speech sounds. *The Journal of the Acoustical Society of America*, 93(4):2160–2171.

Glasberg, B. and Moore, B. (1986). Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. *The Journal of the Acoustical Society of America*, 79(4):1020–1033.

Glasberg, B. and Moore, B. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1):103–138.

Godoy, E. and Stylianou, Y. (2012). Unsupervised acoustic analyses of normal and lombard speech, with spectral envelope transformation to improve intelligibility. In *Proc. Interspeech*, pages 1472–1475, Portland, USA.

Gold, B. and Rader, C. (1967). The channel vocoder. *IEEE Trans. Audio and Electroacoustics*, 15(4):148–161.

Goldsworthy, R. and Greenberg, J. (2004). Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *The Journal of the Acoustical Society of America*, 116(6):3679–3689.

Haas, H. (1972). The influence of a single echo on the audibility of speech. *Journal of the Audio Engineering Society*, 20(2):146–159.

Hall, J. and Flanagan, J. (2010). Intelligibility and listener preference of telephone speech in the presence of babble noise. *The Journal of the Acoustical Society of America*, 127(1):280–285.

Hansen, J. (1996). Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech communication*, 20(1):151–173.

Hazan, V. and Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *The Journal of the Acoustical Society of America*, 130(4):2139–2152.

Helfer, K. (1998). Auditory and auditory-visual recognition of clear and conversational speech by older adults. *Journal American Academy of Audiology*, 9:234–242.

Hilkhuysen, G., Gaubitch, N., Brookes, M., and Huckvale, M. (2014). Effects of noise suppression on intelligibility. II: An attempt to validate physical metricsa). *The Journal of the Acoustical Society of America*, 135(1):439–450.

Hodoshima, N., Arai, T., and Kusumoto, A. (2002). Enhancing temporal dynamics of speech to improve intelligibility in reverberant environments. In *Proc. Forum Acusticum*, Sevilla.

Hodoshima, N., Arai, T., Kusumoto, A., and Kinoshita, K. (2006). Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments. *The Journal of the Acoustical Society of America*, 119(6):4055–4064.

Hohmann, V. (2002). Frequency analysis and synthesis using a gammatone filterbank. *Acta Acustica united with Acustica*, 88(3):433–442.

Holdsworth, J., Nimmo-Smith, I., Patterson, R., and Rice, P. (1988). Implementing a gammatone filter bank. Technical report, MRC Applied Psychology Unit, Cambridge.

Holland, J. (1975). *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor.

Holube, I. and Kollmeier, B. (1996). Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *Journal of the Acoustical Society of America*, 100(3):1703–1716.

Hood, J. and Poole, J. (1980). Influence of the speaker and other factors affecting speech intelligibility. *Audiology*, 19(5):434–455.

Houtgast, T. and Steeneken, H. (1973). The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acta Acustica united with Acustica*, 28(1):66–73.

Houtgast, T. and Steeneken, H. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America*, 77(3):1069–1077.

Houtgast, T., Steeneken, H., and Plomp, R. (1980). Predicting speech intelligibility in rooms from the modulation transfer function. I. general room acoustics. *Acta Acustica united with Acustica*, 46(1):60–72.

Howell, P., Barry, W., and Vinson, D. (2006). Strength of british english accents in altered listening conditions. *Perception & psychophysics*, 68(1):139–153.

Huang, D., Rahardja, S., and Ong, E. (2010). Lombard effect mimicking. In *Proc. of ISCA Workshop on Speech Synthesis*, pages 258–263.

Irino, T. and Patterson, R. (2006). A dynamic compressive gammachirp auditory filterbank. *IEEE Tran. Audio, Speech, and Language Processing*, 14(6):2222–2232.

Jokinen, E., Alku, P., and Vainio, M. (2012). Comparison of post-filtering methods for intelligibility enhancement of telephone speech. In *Proc. of European Signal Processing Conference (EUSIPCO)*, pages 2333–2337.

Jørgensen, S. and Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *The Journal of the Acoustical Society of America*, 130(3):1475–1487.

Junqua, J. (1993). The lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, 93(1):510–524.

Junqua, J. (1996). The influence of acoustics on speech production: A noise-induced stress phenomenon known as the lombard reflex. *Speech Communication*, 20(1):13–22.

Junqua, J., Fincke, S., and Field, K. (1998). Influence of the speaking style and the noise spectral tilt on the lombard reflex and automatic speech recognition. In *Proc. ICSLP*, pages 467–470.

Junqua, J., Fincke, S., and Field, K. (1999). The lombard effect: A reflex to better communicate with others in noise. In *Proc. ICASSP*, pages 2083–2086.

Jürgens, T. and Brand, T. (2009). Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model. *The Journal of the Acoustical Society of America*, 126(5):2635–2648.

Jürgens, T., Brand, T., and Kollmeier, B. (2007). Modelling the human-machine gap in speech reception: microscopic speech intelligibility prediction for normal-hearing subjects with an auditory model. In *Proc. Interspeech*, pages 410–413.

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall, Englewood Cliffs, New Jersey.

Kates, J. and Arehart, K. (2005). Coherence and the speech intelligibility index. *The Journal of the Acoustical Society of America*, 117(4):2224–2237.

Kawahara, H., Masuda-Katsuse, I., and De Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech communication*, 27(3):187–207.

Kitamura, T., Kinoshita, K., Arai, T., Kusumoto, A., and Murahara, Y. (2000). Designing modulation filters for improving speech intelligibility in reverberant environments. In *Proc. Interspeech*, pages 586–589.

Kleijn, W., Crespo, J., Hendriks, R., Petkov, P., Sauert, B., and Vary, P. (2015). Optimizing speech intelligibility in a noisy environment: A unified view. *IEEE Signal Processing Magazine*, 32(2):43–54.

Knudsen, V. (1929). The hearing of speech in auditoriums. *The Journal of the Acoustical Society of America*, 1(1):56–82.

Kondo, K. (2012). Speech quality. In *Subjective Quality Measurement of Speech*, pages 7–20. Springer Berlin Heidelberg.

Koutsogiannaki, M., Petkov, P., and Stylianou, Y. (2015). Intelligibility enhancement of casual speech for reverberant environments inspired by clear speech properties. In *Proc. IEEE Int. Conf. Speech Communication Association*.

Koutsogiannaki, M. and Stylianou, Y. (2014). Simple and artefact-free spectral modifications for enhancing the intelligibility of casual speech. In *Proc. ICASSP*, pages 4648–4652.

Krause, J. and Braida, L. (2002). Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *The Journal of the Acoustical Society of America*, 112(5):2165–2172.

Krause, J. and Braida, L. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America*, 115(1):362–378.

Kretsinger, E. and Young, N. (1960). The use of fast limiting to improve the intelligibility of speech in noise. *Communications Monographs*, 27(1):63–69.

Kryter, K. (1962). Methods for the calculation and use of the articulation index. *The Journal of the Acoustical Society of America*, 34(11):1689–1697.

Kusumoto, A., Arai, T., Kinoshita, K., Hodoshima, N., and Vaughan, N. (2005). Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments. *Speech communication*, 45(2):101–113.

Kusumoto, A., Arai, T., Kitamura, T., Takahashi, M., and Murahara, Y. (2000). Modulation enhancement of speech as a preprocessing for reverberant chambers with the hearing-impaired. In *Proc. ICASSP*, pages 853–856.

Lagarias, J., Reeds, J., Wright, M., and Wright, P. (1998). Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on Optimization*, 9(1):112–147.

Langhans, T. and Strube, H. (1982). Speech enhancement by nonlinear multiband envelope filtering. In *Proc. ICASSP*, pages 156–159.

Langner, B. and Black, A. (2005). Improving the understandability of speech synthesis by modeling speech in noise. In *Proc. ICASSP*, pages 265–268.

Laroche, J. and Dolson, M. (1999). New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects. In *Proc. Applications of Signal Processing to Audio and Acoustics*, pages 91–94.

Li, F., Menon, A., and Allen, J. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *The Journal of the Acoustical Society of America*, 127(4):2599–2610.

Li, F., Trevino, A., Menon, A., and Allen, J. (2012). A psychoacoustic method for studying the necessary and sufficient perceptual cues of american english fricative consonants in noise. *The Journal of the Acoustical Society of America*, 132(4):2663–2675.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle, W. and Marchal, A., editors, *Speech production and speech modelling*, pages 403–439. Kluwer Academic Publishers.

Liu, S., Del Rio, E., Bradlow, A., and Zeng, F. (2004). Clear speech perception in acoustic and electric hearing. *The Journal of the Acoustical Society of America*, 116(4):2374–2383.

Lombard, E. (1911). Le signe de l'elevation de la voix. *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, 37(101-119).

Lu, Y. and Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. *The Journal of the Acoustical Society of America*, 124(5):3261–3275.

Lu, Y. and Cooke, M. (2009a). The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, 51(12):1253–1262.

Lu, Y. and Cooke, M. (2009b). Speech production modifications produced in the presence of low-pass and high-pass filtered noise. *The Journal of the Acoustical Society of America*, 126(3):1495–1499.

Maher, R. (1991). Sinewave additive synthesis revisited. In *Proc. 91st Audio Eng. Soc. Conv.*, New York, NY.

McLoughlin, I. and Chance, R. (1997). LSP-based speech modification for intelligibility enhancement. In *Proc. of International Conference on Digital Signal Processing*, pages 591–594.

Miller, G. and Nicely, P. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, 27(2):338–352.

Moore, B. (1982). *An introduction to the psychology of hearing*, volume 4. Academic press London.

Moore, B. and Glasberg, B. (1996). A revision of Zwicker's loudness model. *Acta Acustica united with Acustica*, 82(2):335–345.

Moore, B., Glasberg, B., and Baer, T. (1997). A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, 45(4):224–240.

Moore, B. C. and Moore, B. (2004). *An introduction to the psychology of hearing*. Academic Press London, 5th edition.

Nábělek, A., Letowski, T., and Tucker, F. (1989). Reverberant overlap-and self-masking in consonant identification. *The Journal of the Acoustical Society of America*, 86(4):1259–1265.

Nelder, J. and Mead, R. (1965). A simplex method for function minimization. *The computer journal*, 7(4):308–313.

Niederjohn, R. and Grotelueschen, J. (1976). The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. *IEEE Tran. Acoustics, Speech and Signal Processing*, 24(4):277–282.

Niederjohn, R. and Grotelueschen, J. (1978). Speech intelligibility enhancement in a power generating noise environment. *IEEE Tran. Acoustics, Speech and Signal Processing*, 26(4):378–380.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Park, H., Yoon, J., Kim, J., and Oh, E. (2010). Improving perceptual quality of speech in a noisy environment by enhancing temporal envelope and pitch. *IEEE Signal Processing Lett.*, 17(5):489–492.

Patel, R. and Schell, K. (2008). The influence of linguistic content on the lombard effect. *Journal of Speech, Language, and Hearing Research*, 51(1):209–220.

Patterson, R., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1988). An efficient auditory filterbank based on the gammatone function. APU report 2341, Applied Psychology Unit, Cambridge.

Payton, K., Uchanski, R., and Braida, L. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America*, 95(3):1581–1592.

Petkov, P., Henter, G., and Kleijn, W. (2013). Maximizing phoneme recognition accuracy for enhanced speech intelligibility in noise. *IEEE Trans. Audio, Speech, Lang. Processing*, 21(5):1035–1045.

Phatak, S., Lovitt, A., and Allen, J. (2008). Consonant confusions in white noise. *The Journal of the Acoustical Society of America*, 124(2):1220–1233.

Picheny, M., Durlach, N., and Braida, L. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech, Language and Hearing Research*, 28(1):96–103.

Picheny, M., Durlach, N., and Braida, L. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language and Hearing Research*, 29(4):434–446.

Picheny, M., Durlach, N., and Braida, L. (1989). Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. *Journal of Speech, Language and Hearing Research*, 32(3):600–603.

Pittman, A. and Wiley, T. (2001). Recognition of speech produced in noise. *Journal of Speech, Language, and Hearing Research*, 44(3):487–496.

Raitio, T., Suni, A., Vainio, M., and Alku, P. (2011). Analysis of HMM-based lombard speech synthesis. In *Proc. Interspeech*, pages 2781–2784.

Raj, B., Seltzer, M., and Stern, R. (2004). Reconstruction of missing features for robust speech recognition. *Speech communication*, 43(4):275–296.

Rasetshwane, D., Boston, J., Li, C., Durrant, J., and Genna, G. (2009). Enhancement of speech intelligibility using transients extracted by wavelet packets. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 173–176.

Régnier, M. and Allen, J. (2008). A method to identify noise-robust perceptual features: Application for consonant/t. *The Journal of the Acoustical Society of America*, 123(5):2801–2814.

Rhebergen, K. and Versfeld, N. (2005). A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 117(4):2181–2192.

Rhebergen, K., Versfeld, N., and Dreschler, W. (2005). Release from informational masking by time reversal of native and non-native interfering speech. *The Journal of the Acoustical Society of America*, 118(3):1274–1277.

Sabin, W. and Schoenike, E. (1998). *HF Radio Systems & Circuits*. Noble Publishing Corporation.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Tran. Acoustics, Speech and Signal Processing*, 26(1):43–49.

Sauert, B., Enzner, G., and Vary, P. (2006). Near end listening enhancement with strict loudspeaker output power constraining. In *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*.

Sauert, B., Löllmann, H., and Vary, P. (2008). Near end listening enhancement by means ofwarped low delay filter-banks. In *Proc. ITG Conference on Speech Communication*, pages 1–4.

Sauert, B. and Vary, P. (2006a). Improving speech intelligibility in noisy environments by near end listening enhancement. In *Proc. ITG Conference on Speech Communication*.

Sauert, B. and Vary, P. (2006b). Near end listening enhancement: Speech intelligibility improvement in noisy environments. In *Proc. ICASSP*, pages 493–496.

Sauert, B. and Vary, P. (2009). Near end listening enhancement optimized with respect to speech intelligibility index. In *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, volume 17, pages 1844–1848.

Sauert, B. and Vary, P. (2010a). Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations. In *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pages 1919–1923.

Sauert, B. and Vary, P. (2010b). Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement. In *Proc. of ITG-Fachtagung Sprachkommunikation*.

Sauert, B. and Vary, P. (2012). Near-end listening enhancement in the presence of bandpass noises. In *Proc. ITG Conference on Speech Communication*, pages 195–198.

Schwartz, R. and Austin, S. (1990). Efficient, high-performance algorithms for N-best search. In *Proc. Speech and Natural Language Workshop*, pages 6–11.

Shin, J., Jin, Y., Park, S., and Kim, N. (2009). Speech reinforcement based on partial masking effect. In *Proc. ICASSP*, pages 4401–4404.

Shin, J., Lim, W., Sung, J., and Kim, N. (2007). Speech reinforcement based on partial specific loudness. In *Proc. Interspeech*, pages 978–981.

Simpson, S. and Cooke, M. (2005). Consonant identification in N-talker babble is a nonmonotonic function of N. *The Journal of the Acoustical Society of America*, 118(5):2775–2778.

Singh, R. and Allen, J. (2012). The influence of stop consonants' perceptual features on the articulation index model. *The Journal of the Acoustical Society of America*, 131(4):3051–3068.

Skowronski, M. and Harris, J. (2006). Applied principles of clear and lombard speech for automated intelligibility enhancement in noisy environments. *Speech Communication*, 48(5):549–558.

Smiljanić, R. and Bradlow, A. (2007). Clear speech intelligibility: Listener and talker effects. In *Proc. ICPHS*, pages 661–664, Saarbrucken, Germany.

Smiljanić, R. and Bradlow, A. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and linguistics compass*, 3(1):236–264.

Steeneken, H. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, 67(1):318–326.

Taal, C., Hendriks, R., and Heusdens, R. (2012a). A low-complexity spectro-temporal distortion measure for audio processing applications. *IEEE Tran. Audio, Speech, and Language Processing*, 20(5):1553–1564.

Taal, C., Hendriks, R., and Heusdens, R. (2012b). A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure. In *Proc. ICASSP*, pages 4061–4064, Portland Oregon, USA.

Taal, C., Hendriks, R., Heusdens, R., and Jensen, J. (2010). On predicting the difference in intelligibility before and after single-channel noise reduction. In *Proc. Int. Workshop, Acoust. Echo Noise Control*.

Taal, C., Hendriks, R., Heusdens, R., and Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Tran. Audio, Speech, and Language Processing*, 19(7):2125–2136.

Taal, C. and Jensen, J. (2013). Sii-based speech preprocessing for intelligibility improvement in noise. In *Interspeech*, pages 3582–3586.

Taal, C., Jensen, J., and Leijon, A. (2013). On optimal linear filtering of speech for near-end listening enhancement. *IEEE Signal Processing Lett.*, 20(3):225–228.

Taal, C. H., Hendriks, R. C., and Heusdens, R. (2014). Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure. *Computer Speech & Language*, 28(4):858–872.

Tang, Y. and Cooke, M. (2010). Energy reallocation strategies for speech enhancement in known noise conditions. In *Proc. Interspeech*, pages 1636–1639.

Tang, Y. and Cooke, M. (2011). Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. In *Proc. Interspeech*, pages 345–348, Florence, Italy.

Tang, Y. and Cooke, M. (2012). Optimised spectral weightings for noise-dependent speech intelligibility enhancement. In *Proc. Interspeech*, Portland Oregon, USA.

Tantibundhit, C., Boston, J., Li, C., Durrant, J., Shaiman, S., Kovacyk, K., and El-Jaroudi, A. (2007). New signal decomposition method based speech enhancement. *Signal Processing*, 87(11):2607–2628.

Thomas, I. (1968). The influence of first and second formants on the intelligibility of clipped speech. *Journal of the Audio Engineering Society*, 16(2):182–185.

Thomas, I. and Niederjohn, R. (1968). Enhancement of speech intelligibility at high noise levels by filtering and clipping. *Journal of the Audio Engineering Society*, 16(4):412–415.

Thomas, I. and Niederjohn, R. (1970). The intelligibility of filtered-clipped speech in noise. *Journal of the Audio Engineering Society*, 18(3):299–303.

Thomas, I. and Ohley, W. (1972). Intelligibility enhancement through spectral weighting. In *Proc. of Conf. on Speech Communication and Processing*, pages 360–363.

Toscano, J. and Allen, J. (2014). Across-and within-consonant errors for isolated syllables in noise. *Journal of Speech, Language, and Hearing Research*, 57(6):2293–2307.

Uchanski, R. (2005). Clear speech. In Pisoni, D. and Remez, R., editors, *The handbook of speech perception*, pages 207–235. Blackwell, Malden, MA.

Uther, M., Knoll, M., and Burnham, D. (2007). Do you speak E-NG-LI-SH? a comparison of foreigner-and infant-directed speech. *Speech Communication*, 49(1):2–7.

Valentini-Botinhao, C., Maia, R., Yamagishi, J., King, S., and Zen, H. (2012). Cepstral analysis based on the glimpse proportion measure for improving the intelligibility of HMM-based synthetic speech in noise. In *Proc. ICASSP*, pages 3997–4000.

Valentini-Botinhao, C., Yamagishi, J., King, S., and Stylianou, Y. (2013). Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of HMM-based synthetic speech in noise. In *Proc. Interspeech*, pages 3567–3571.

Van Summers, W., Pisoni, D., Bernacki, R., Pedlow, R., and Stokes, M. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, 84(3):917–928.

Wang, M. and Bilger, R. (1973). Consonant confusions in noise: A study of perceptual features. *The Journal of the Acoustical Society of America*, 54(5):1248–1266.

Watkins, A. (2005). Perceptual compensation for effects of reverberation in speech identification. *The Journal of the Acoustical Society of America*, 118(1):249–262.

Webster, J. and Klumpp, R. (1962). Effects of ambient noise and nearby talkers on a face-to-face communication task. *The Journal of the Acoustical Society of America*, 34(7):936–941.

Weintraub, M. (1985). *A theory and computational model of auditory monaural sound separation*. PhD thesis, Stanford University.

Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A., and Kollmeier, B. (2005). Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines. In *Proc. Interspeech*, pages 1273–1276.

Womack, B. and Hansen, J. (1996). Classification of speech under stress using target driven features. *Speech Communication*, 20(1):131–150.

Yoo, S., Boston, J., El-Jaroudi, A., Li, C., Durrant, J., Kovacyk, K., and Shaiman, S. (2007). Speech signal modification to increase intelligibility in noisy environments. *The Journal of the Acoustical Society of America*, 122(2):1138–1149.

Zaar, J. and Dau, T. (2015). Sources of variability in consonant perception of normal-hearing listeners. *The Journal of the Acoustical Society of America*, 138(3):1253–1267.

Zorila, T.-C., Kandia, V., and Stylianou, Y. (2012). Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In *Proc. Interspeech*, Portland, USA.

Zorila, T.-C. and Stylianou, Y. (2014). On spectral and time domain energy reallocation for speech-in-noise intelligibility enhancement. In *Proc. Interspeech*, pages 2050–2054.

Appendix A

Parameter set of the gammatone fiterbank analysis-resynthesis design

Table A.1 shows the parameter set of the gammatone filterbank design and the corresponding resynthesis design, used throughout this thesis. The filterbank is designed for 25 kHz sampling frequency, 50Hz is the lower cutoff frequency, 8 kHz is the upper cutoff frequency, and finally, filters per ERB set to 1.

Table A.1: Parameter set of the fiterbank analysis-resynthesis design. The table gives the centre frequency f_c in Hz. For the analysis part, the filter coefficient $\tilde{a}_f$ is given. For the resynthesis part, the delay in samples Δt_f, the phase factor $\tilde{b}_f$ and the gain g_f in dB are given.

		Analysis part	Resynthesis part		
Channel Num	f_c	$\tilde{a}_f$	Δt_f	$\tilde{b}_f$	$20\log_{10}(g_f)$
1	50.000000	0.992247 + 0.012470i	0	0.977897 + -0.209087i	0.47
2	81.780913	0.991244 + 0.020377i	0	0.302571 + -0.953127i	1.84
3	117.183999	0.990055 + 0.029167i	0	-0.704287 + -0.709916i	1.94
4	156.622087	0.988639 + 0.038936i	0	-0.939312 + 0.343064i	1.62
5	200.555061	0.986950 + 0.049789i	0	-0.006646 + 0.999978i	1.08
6	249.495215	0.984930 + 0.061841i	0	0.965057 + 0.262041i	0.41
7	304.013236	0.982510 + 0.075217i	0	0.392069 + -0.919936i	-0.33
8	364.744850	0.979602 + 0.090053i	0	-0.921700 + -0.387904i	-1.11
9	432.398240	0.976103 + 0.106496i	0	-0.227982 + 0.973665i	-1.86
10	507.762304	0.971885 + 0.124704i	0	0.992249 + -0.124264i	-2.51
11	591.715851	0.966792 + 0.144845i	0	-0.636002 + -0.771687i	-3.00
12	685.237851	0.960635 + 0.167095i	0	-0.080664 + 0.996741i	-3.44
13	789.418854	0.953184 + 0.191635i	0	0.553411 + -0.832908i	-3.66
14	905.473698	0.944157 + 0.218650i	6	-0.825815 + -0.563942i	-3.71
15	1034.755684	0.933216 + 0.248318i	16	-0.989389 + -0.145288i	-3.61
16	1178.772351	0.919950 + 0.280806i	25	-0.974133 + 0.225976i	-3.59
17	1339.203058	0.903868 + 0.316256i	33	-0.847708 + 0.530463i	-3.59
18	1517.918566	0.884378 + 0.354765i	40	-0.623029 + 0.782199i	-3.61
19	1717.002852	0.860780 + 0.396367i	46	-0.256063 + 0.966660i	-3.59
20	1938.777411	0.832246 + 0.440994i	52	-0.172465 + 0.985016i	-3.66
21	2185.828325	0.797814 + 0.488440i	57	0.060470 + 0.998170i	-3.58
22	2461.036421	0.756376 + 0.538302i	62	-0.058942 + 0.998261i	-3.55
23	2767.610865	0.706684 + 0.589908i	66	0.086823 + 0.996224i	-3.66
24	3109.126578	0.647374 + 0.642229i	69	0.615464 + 0.788165i	-3.61
25	3489.565929	0.577007 + 0.693769i	73	0.116736 + 0.993163i	-3.56
26	3913.365169	0.494161 + 0.742434i	76	0.042171 + 0.999110i	-3.63
27	4385.466163	0.397581 + 0.785398i	78	0.634485 + 0.772935i	-3.60
28	4911.374017	0.286413 + 0.818957i	81	-0.109235 + 0.994016i	-3.68
29	5497.221272	0.160546 + 0.838433i	83	-0.074657 + 0.997209i	-3.51
30	6149.839410	0.021103 + 0.838144i	85	-0.368707 + 0.929546i	-3.60
31	6876.838525	-0.128923 + 0.811559i	87	-0.887930 + 0.459978i	-4.35
32	7686.696054	-0.283884 + 0.751735i	88	-0.371543 + 0.928416i	-1.24