2020-04-29

New Approaches to Measuring and Improving Speech Intelligibility

Tanaka, Mai

UNIVERSITY OF CALGARY

New Approaches to Measuring and Improving Speech Intelligibility

by

Mai Tanaka

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN ELECTRICAL ENGINEERING

CALGARY, ALBERTA

APRIL, 2020

Abstract

Hearing impaired and elderly listeners have trouble understanding speech in the presence of noise, which may be caused by the reduced speech processing ability in the brain. Slowing down the speech that is presented, may provide the listener's brain more opportunities to process stimuli or collect crucial speech cues. Speech rate modulation algorithms that generate a slowed down version of speech have been suggested as an approach to improve comprehension. However, existing non-uniform speech rate modulation algorithms are difficult to experimentally replicate, often requiring specialized linguistics knowledge of the speech, such as the locations of phonetic elements. Additionally, no previous study has objectively measured intelligibility of rate modulated speech. This thesis proposes new methods for objectively measuring rate modulated speech characteristics through Speech Rate Based Intelligibility Metrics (SRBIM), and for improving speech through SOLely Acoustic Non-Uniform (SOLANU) speech rate modulation. SRBIM was compared to existing metrics in situations where the existing metrics were known to work, involving speech contaminated by various noise types and noise levels, and speech processed through three types of noise reduction algorithms. New and existing metric behaviours were compared by analyzing the correlation and scatter plots between all possible metric pairs. The metrics were also used to rank the performances of the three noise reduction algorithms. Simulations were extended to speech processed through two existing methods and two variants of the proposed SOLANU speech rate modulation algorithm. Analysis through correlation and scatter plots, as well as subjective observations were included. When compared against two of the three existing metrics, SRBIM showed a linear relationship in measuring speech intelligibility of noisy speech and speech processed through noise reduction. Subjective observations showed that clean rate modulated speech was more intelligible than noisy rate modulated speech. SRBIM were the only metrics that reflected the impact of rate modulation on clean speech compared to noisy speech. Thus, SRBIM has potential in measuring speech intelligibility of noisy speech, and speech that had been enhanced through noise reduction or rate modulation. Future research involving human observers with rate modulated speech is recommended to further validate SRBIM to subjective intelligibility scores.

Preface

This thesis is original, unpublished, independent work by the author, M. Tanaka.

Acknowledgments

I'd like to express my appreciation towards all those who have made the completion of this thesis possible. Especially Dr. Smith, my supervisor. Your invaluable advice on conducting research, writing up results, and presentation have made me a much better researcher. I'd also like to thank members of my examining committee, Dr. Yanushkevich, Dr. Sesay and Dr. Curiel, for sitting in on my thesis defence and expressing interest to my research. Considering the challenges that COVID-19 has presented to the completion of the thesis, I would not have been able to pull through if it were not for the complete support from all of you.

I'd also like to thank my fellow graduate students in my laboratory. Special thanks to Ishani for letting me bounce ideas back and forth, and the odd fun talks that would reinvigorate me to carry on with my work. It was great to have a "sister-in-arms" that I could share the ups and downs of my research.

I'd also like to acknowledge the agencies which has funded me throughout my graduate years and given me one less thing to worry about. Scholarships that I received from the National Sciences and Engineering Research Council of Canada, the Campbell McLaurin Foundation for Hearing Deficiencies, and the Government of Alberta have assisted me in completing my thesis.

I'd also like to extend my thanks to the administrative staff who have supported me throughout my graduate studies. I have e-mailed several members of the team with urgent requests, and approached the office a minute before closing time several times. Every time the staff has handled my situation supportively and with a smile. If not for you, my graduate experience would not have been as pleasant as it was.

Last but not least, I'd like to thank my family for their full support throughout my post-secondary education. My mother had shown great understanding throughout my endeavour, which I greatly appreciate. My father gave me another perspective into my thesis which has been extremely beneficial. Kulolo, our dog whom passed away in 2018, gave me mental support countless times during my graduate education.

Through the completion of this thesis, I have found that investigations set out to solve a particular research question inevitably leads to new research questions. Setting out to solve these new research questions, branches into even more research questions. In that regard, research is an iterative process with no end goal. Perhaps, the love of research is an addiction to the tiny triumphs along this cyclical journey, which is something I can relate to. While there were weeks,

sometimes months at a time when I had felt frustrated or lost, the excitement that I experienced when something did go right, or when I felt I was “onto something”, was a feeling that would be difficult to experience elsewhere. However, for the purposes of submitting a completed thesis, an adequate “break” in the cyclical process must be identified, followed by in-depth reflections and analysis, and the presentation of the findings “thus far” to peers in a logical fashion. My hopes, as I conclude this this thesis, are that I have adequately fulfilled these objectives, and that my journey into research will continue for years to come.

Table of Contents

List of Tables

List of Figures

List of Symbols, Abbreviations and Nomenclature

AI	Articulation Index
AMDF	Average Magnitude Difference Function
CDC	Centers for Disease Control and Prevention
CSII	Coherence Speech Intelligibility Index
ESOLA	Epoch Synchronous Overlap and Add
ESTOI	Extended Short-Time Objective Intelligibility
FD-PSOLA	Frequency Domain Pitch Synchronous Overlap and Add
HMM	Hidden Markov Model
HNM	Harmonic plus Noise Model
HTK	Hidden Markov Model Toolkit
LACE^{TM}	Listening and Communication Enhancement
LMPSC	Log Mel Power Spectral Coefficient
MFCC	Mel Frequency Cepstral Coefficient
NCHS	National Center for Health Statistics
nsec	Normalized Subband Envelope Correlation
OLA	Overlap and Add
PSOLA	Pitch Synchronous Overlap and Add
sEPSM	speech-based Envelope Power Spectrum Model
SII	Speech Intelligibility Index
SNR	Signal to Noise Ratio
SOLA	Synchronous Overlap and Add
SOLANU	SOLely Acoustic Non-Uniform (speech rate modulation)
SRBIM	Speech Rate Based Intelligibility Metric
SRT	Speech Reception Threshold
STI	Speech Transmission Index
STM	Spectral Transition Measure
STOI	Short-Time Objective Intelligibility
TD-PSOLA	Time Domain Pitch Synchronous Overlap and Add
WSOLA	Waveform Similarity Overlap and Add

1 Introduction

Chapter 1 is an overview of the general problems involved with individuals trying to understand speech in the presence of noise, and a discussion of the main focus, the value and the objectives of this thesis in solving this problem. The chapter starts with a background that motivated this thesis, including the innate ability of the average human to extract relevant speech from noisy environments. However, certain populations struggle to comprehend noisy speech, and current assistive technologies are proving inadequate in improving speech intelligibility. In this thesis, the term speech intelligibility refers to the level of understanding achieved through the presented speech. It is a characteristic of the speech which ultimately results in speech comprehension, which requires the brain as well as the ear to be working together.

Presentation of speech at a slower rate may be able to accommodate for the slower brain, which is part of the cause of hearing difficulties in noise in elderly and hearing impaired listeners. The decreased processing ability has been supported in neural studies where an increased delay of the brains' responses to sounds have been observed in older subjects compared to responses in younger subjects (Tremblay, Piskosz, & Souza, 2003). As such, speech enhancement algorithms that automatically produce a slowed down version of speech through rate modulation may be beneficial in improving comprehension, as this gives the listener's brain adequate time to process the speech.

However, there have been very little research on non-uniform speech rate modulation algorithms, or algorithms that are capable of producing an output speech with a different speech rate pattern compared to the input speech, and objective speech intelligibility metrics that measure improvements due to the associated time-scale modifications. Furthermore, this thesis will validate the newly proposed algorithms against existing speech intelligibility metrics with simply noisy speech, speech enhanced through noise reduction, and speech that has been rate modulated.

The literature on speech rate modulation uses a variety of different terminologies when discussing speech stretched in time of speech that has been slowed down. The most common name is time scaled speech or time scale modification. Other papers refer to the technique as time stretching. This thesis will refer to the same process as speech rate modulation for better alignment with the newly proposed speech intelligibility metrics, which are based on speech rate.

The resulting speech may be referred to as rate modulated, time scaled or stretched speech, depending on the given context.

Following the background information, there is a description of the research focus and the value of the research. Hearing impairments are speculated to be a costly health issue in the future, that should not be ignored especially with a longer life expectancy. The overall research aim and individual objectives fulfilled in this thesis will also be clearly stated. The last sections elucidate the thesis outline and the student's contributions.

1.1 Background

While many normal hearing people listen to and understand speech effortlessly in noise, there are specific populations who suffer greatly from the effects of noise. Slowing speech down by changing its speech rate, seems to improve intelligibility for these noise vulnerable listeners. The thesis focuses on methods of measuring speech intelligibility using knowledge of speech rate, as well as approaches of improving intelligibility by modulating speech rate.

1.1.1 The Cocktail Party Phenomenon

Rarely do humans converse in the complete absence of background noise. Whether it be the sound of the car tires hitting the pavement, music playing inside a store or other people busily and simultaneously talking in a large cafeteria, verbal communication amidst noise is a common occurrence.

The innate ability of humans to decipher speech in noise is a well-studied phenomenon. This effect was formally discovered during a study investigating speech recognition in listeners simultaneously presented with two different auditory stimuli, and has been labeled as the cocktail party phenomenon (Cherry, 1953). The phenomenon refers to the ability of people to identify, segregate and track a single sound source, while suppressing all other unwanted noise. Most normal hearing individuals are skilled at this task, enabling them to understand speech even in the presence of noise. However, there are specific populations that have difficulty comprehending speech amidst noise.

1.1.2 Noise Vulnerable Populations

The cocktail party phenomenon allows most people to comprehend noisy speech. Nonetheless, there are certain subsets of the population that cannot execute this skill and suffer from subpar speech comprehension levels in the presence of background noise. Namely, people with hearing impairments and elderly populations regularly exhibit a vulnerability to the effects

of noise on speech. Several studies have reported significantly lower speech recognition performance in hearing impaired or elderly listeners in comparison to normal hearing or young subjects. For example, Speech Reception Threshold (SRT), or the relative presentation loudness of speech to noise required for 50 % intelligibility in a listener, was worse for both young and old listeners with hearing impairments compared to young normal hearing subjects (Peters, Moore, & Baer, 1998). Similarly, speech recognition scores were significantly worse in hearing impaired subjects who were unable to take advantage of the transient alleviations in the presented noise (Bacon, Opie, & Montoya, 1998). In a study investigating the relationship between speech rates and comprehension, elderly listeners had lower speech recognition in both quiet and in noise with time compressed speech, or speech that had been sped up, when compared to younger listeners (Gordon-Salant, Fitzgibbons, & Yeni-Komshian, 2011; Gordon-Salant & Friedman, 2011). Furthermore, this detrimental condition is not a rarity.

An extensive 2009 National Health Interview Survey conducted by the Centers for Disease Control and Prevention's (CDC) National Center for Health Statistics (NCHS) in the United States reported that 15 % of all adults over the age of 18 experienced some difficulty in hearing without a hearing aid (Pleis, Ward, & Lucas, 2010). As seen in Table 1-1, the prevalence of hearing impairments increases with age (Pleis et al., 2010), as one of the most common forms of hearing impairments is age related hearing loss. With the aging society, there are projections that adult onset hearing loss will be among the top ten causes of disease burden in high- and middle-income countries by 2030 (Mathers & Loncar, 2006).

Hearing impairments are also debilitating because the primary problem of not being able to understand conversations in everyday settings can lead to secondary complications. For example, in seniors, hearing impairments have been linked with cognitive decline, cognitive impairments, dementia, social isolation and increased mortality rates (Lesica, 2018; Loughrey, Kelly, Kelley, Brennan, & Lawlor, 2018). In one study, adults aged 70 to 79 years with hearing loss had a 30 % to 40 % accelerated rate of cognitive decline and 24 % higher risk in developing

Table 1-1: Selected Data on Frequency of Hearing Trouble as per Age Group, United States, 2009 [Adapted from (Pleis et al., 2010)]

Age group	Total number	Number with hearing trouble	Percentage
18-44 years	110337	7059	6.4 %
45-64 years	79195	13903	17.6 %
65-74 years	20597	5840	28.4 %
75 years and over	17242	7686	44.6 %
All	227371	34488	15.2 %

cognitive impairment as compared to normal hearing individuals over a span of six years (Lin et al., 2013). In women aged 60 to 69 years, greater hearing loss has been linked with social isolation (Mick, Kawachi, & Lin, 2014). Hence, difficulties in understanding speech in noise affects an unignorably significant number of people and keeping the problem unattended can lead to further issues.

1.1.3 Current Extent of Assistive Hearing Technologies

The current remedy in solving the problem of hearing difficulties in noise is to use hearing devices. This can be as drastic as undergoing cochlear implant surgery, or more common measures such as being fitted with hearing aids. Indeed, using hearing aids does seem to mitigate the effects of hearing impairments on cognitive issues. In an American cohort, it was found that the decline in cognitive outcomes, measured through episodic memory scores, was slowed down by using hearing aids (Maharani, Dawes, Nazroo, Tampubolon, & Pendleton, 2018).

However, the proportion of people who actually use hearing aids is low in comparison to the number of people who would benefit from use. This observation has been documented across different countries. In an Australian cohort involving 1371 participants over the age of 49, it was found that 318 out of 1224 (26.0 %), tested as having hearing loss, but did not even own a hearing aid (Gopinath et al., 2011). Additionally, 35 out of 138 (26.4 %) individuals had hearing loss and owned hearing aids, but did not use their devices (Gopinath et al., 2011). From these numbers, it can be calculated that 353 out of 456 (77.4 %) seniors with hearing loss either did not own a hearing aid, or owned one, but did not use them at all. In another survey conducted in New Zealand, 40.7 % of the 123 participants with hearing impairments reported as not using hearing aids on a regular basis, even though 16 % of those non-users owned a device (Kelly-Campbell & Lessoway, 2015). Similar findings were reported in a study involving 249 hearing aid owners in an industrialized urban community in Finland, where 30.9 % of the seniors seldom or never used their devices (Salonen et al., 2013). It is clear from these surveys that a large proportion of people with hearing impairments, and thus in need of hearing assistance, opt not to use hearing aids, even if they are in possession of one.

There are a few common reasons given pertaining to the reluctance of hearing aid owners in using their devices. A major determinant to hearing aid uptake and use is stigma (David & Werner, 2016). Many people fear that wearing hearing aids would make them look old and inept and are embarrassed and worried that this image will damage their social relationships

(Hindhede, 2011). However, the biggest reason for hearing aid owners not using their devices is that current assistive technologies have very little benefit when it comes to improving speech perception in noise (Duquesnoy & Plomp, 1983). A common complaint with hearing aids is that the noisy speech is loud yet unclear (Tremblay, 2015), heard but not understood (Lesica, 2018). The top two reasons, given in the Australian study, by seniors who were recommended a hearing aid, tried it, but did not continue using it, were that it did not help, and that it was uncomfortable (Gopinath et al., 2011). In the Finnish study, the top two problems experienced by hearing aid users were background noise, reported to be a problem in 68.7 % of users, and feedback problems from the hearing aids, found to be a problem in 19.8 % of users (Salonen et al., 2013).

These complaints demonstrate the highly prevalent problem of difficulties comprehending speech in noise for those with hearing impairments. These problems persist, even with the advances in hearing technologies, as seen in the recent study that compared speech recognition scores of hearing impaired and normal hearing individuals in sixteen different types of noisy and reverberant, or echoing, conditions (Xia, Xu, Pentony, Xu, & Swaminathan, 2018). As such, people who use assistive hearing devices may have difficulty understanding noisy speech because these technologies cannot perform at a level equivalent to the biological components associated with sensing sound. Another possible reason for the persisting problem is that while the hearing technologies aim to make up for shortcomings in the ear, a major cause of the reduced speech comprehension in noise may be the inability of the brain in processing speech, which devices such as hearing aids cannot make up for.

1.1.4 The Brain Factor in Speech Comprehension

Assistive devices do not perform as well as the human ear in discriminating sounds. However, the shortcoming of the peripheral components of hearing is not the only contributor to decreased speech comprehension in noise vulnerable populations. Speech comprehension requires the combined efforts of the ear for hearing, and the brain for analyzing. As such, the ability to hear sounds is a prerequisite, but not an exclusive deciding factor to an individual's capacity in understanding speech (Beck, Larsen, & Bush, 2018; Lesica, 2018). In fact, a survey of 2783 participants revealed that 12.0 % of individuals, tested as having normal hearing through audiometric threshold tests, reported of difficulties understanding speech in noise or with conversations comprising of multiple speakers (Tremblay et al., 2015). This supports the

importance of the brain in understanding noisy speech, which audiometric tests do not currently gauge.

The brain's decreased temporal sampling ability with age (Gordon-Salant et al., 2011) is one such shortcoming that can contribute to decreased noisy speech comprehension. The decreased ability to pick up relevant speech cues especially when the noise fluctuates may have different causes in different subsets of the noise vulnerable populations, but may have a universal solution, naturally implemented by humans. Namely, the presentation of speech at a slower rate may benefit elderly or hearing impaired listeners, who suffer from decreased comprehension due to shortcomings in the brain.

In elderly listeners, the aging brain is most often the cause of the reduced ability to understand speech masked by noise. One characteristic of an older brain is a lower temporal resolution in auditory processing. In other words, older adults typically lose precision in the frequency at which they are able to sample the auditory stimuli, resulting in the loss of details. A lower temporal resolution can also mean that there are fewer opportunities for older adults to catch transient, but crucial cues in the speech patterns, especially when there are fewer opportunities in the first place due to the masking noise. When tested with tones containing gaps of silences of different lengths, older adults were less able to accurately discriminate durational changes in these silences compared to younger listeners (Gordon-Salant et al., 2011). In a study comparing the neural responses to auditory stimuli in hearing impaired and normal hearing older subjects against younger normal hearing subjects, it was found that regardless of hearing ability, older adults had an increased latency in the neural response to auditory stimuli in comparison to the younger subjects (Tremblay, Piskosz, & Souza, 2003). Thus, the aging brain has a lower temporal resolution and is slower to respond to auditory stimuli, making it difficult for older adults to comprehend speech in noise.

Hearing impaired listeners also have difficulty in picking out crucial speech cues in fluctuating noise (Jin & Nelson, 2006). This phenomenon is often called "glimpsing" the speech through noise in literature (Jin & Nelson, 2010; Shamma, Elhilali, & Micheyl, 2011). Specifically, a study tested for speech perception in noise, manipulating the noise masker to have variability in its temporal and spectral characteristics, effectively providing glimpses of the target speech through these windows or dips in noise (Peters et al., 1998). The hearing aid users were then properly fitted with their devices with appropriate frequency-response amplification,

effectively matching the hearing capabilities of the normal hearing individuals (Peters et al., 1998). However, hearing impaired subjects still required a higher signal to noise level to match a normal hearing individuals' performance in temporally and spectrally modulated speech patterns (Peters et al., 1998).

In a similar study, hearing impaired listeners underwent amplification such that they showed equal speech recognition performance to normal hearing subjects in quiet and in steady noise (Jin & Nelson, 2006). When the competing noise was deliberately modified to be temporally gated, the hearing impaired individuals performed significantly worse in comparison to control, even under the same amplification that yielded comparable speech recognition performances in quiet and in steady noise (Jin & Nelson, 2006). Specifically, at -5 dB signal to noise ratio (SNR) where the noise level exceeds those of the target speaker, normal hearing listeners were able to recognize 81 % to 97 % of the key words, while hearing impaired subjects were at 36 % to 74 % (Jin & Nelson, 2006). When the listening conditions were further worsened to -10 dB signal to noise ratio, the performances ranged from 64 % to 96 % in normal hearing and 13 % to 58 % in the hearing impaired listeners (Jin & Nelson, 2006). While a later study revealed that low to mid frequency sensitivity in hearing impaired listeners accounted for the varied levels of glimpsing ability (Jin & Nelson, 2010), a temporal solution achieved by changing the speech rate of what is heard, may be effective in alleviating these difficulties, as speech rate is intricately related with intelligibility.

Yet another study tested speech comprehension where the noise mask was released, or absent, at certain time points, and found that hearing impaired listeners were less able to pick up the relevant speech cues during these dips in noise (Bacon et al., 1998). Furthermore, the difference in performance as compared to normal hearing individuals could not be completely explained by the reduced audibility of the audio signals (Bacon et al., 1998). These studies indicate that while normal hearing subjects were able to take advantage of the speech cues observed through the spectral and temporal dips in the background noise, hearing impaired individuals were less able to make use of these same elements due to lower spectral or temporal resolution, or poorer glimpsing ability.

1.1.5 Speech Enhancement by Speech Rate Modulation

Elderly listeners and individuals with hearing impairments are less adept at picking out relevant speech cues amidst noise due to a lower temporal or spectral resolution. Slowing speech

down is one possible solution. In fact, people naturally and unconsciously slow down speech when speaking in noise (Brumm & Zollinger, 2011). Thus, one viable method of enhancing speech is by modulating its rate in a manner similar to the styles adopted by humans.

1.2 Research Focus and Value of Research

Considering the prevalence rate and dire consequences, there is a great need to improve noisy speech intelligibility for people with hearing impairments. With people living longer, higher quality of life can be achieved when members actively participate well into their senior years. Conversely, inaction to improve the common problem of conversing in noise is predicted to be costly (A. Davis et al., 2016). Appropriate speech rate modulation may be able to improve speech intelligibility for these noise vulnerable populations by alleviating the issues in speech comprehension caused by the brain.

The literature supports the view that slowing speech down improves comprehension for hearing impaired and elderly listeners. Keyword identification scores for normal hearing and hearing impaired listeners in noisy conditions were higher when subjects were provided with clear, rather than conversational speech (Payton, Uchanski, & Braida, 1994). This clear speech involved speakers who were asked to pretend that they were speaking in an environment where communication was difficult (Payton et al., 1994), supporting the claim that speech rate modulation is the natural accommodation to improve comprehension. Similarly, in fluctuating noise, normal hearing subjects improved keyword identification scores with retimed and time stretched speech (Cooke & Aubanel, 2017). Even with the absence of noise, slowing speech rate down improved intelligibility for cochlear implant users (Iwasaki, Ocho, Nagura, & Hoshino, 2002), elderly listeners with normal hearing (Cho, Yu, Chun, Seo, & Han, 2014), elderly listeners with hearing impairments (Gordon-Salant, Fitzgibbons, & Friedman, 2007; Kupryjanow & Czyzewski, 2012b), and children, irrespective of their hearing status (Kupryjanow & Czyzewski, 2012b). These studies reflect the potential of speech rate modulation as a method to improve intelligibility for people who are elderly or have a hearing impairment, especially under noisy conditions.

In order to slow speech down, speech rate modulation algorithms are required. While many papers report of uniform speech rate modulation algorithms that alter rate by a constant stretch factor, there are very few studies on non-uniform speech rate modulation. This may seem odd given that the natural tendency of humans is to slow speech down non-uniformly, where

speech is relatively quick in certain components, but slowed down in other segments. Additionally, Lombard speech, the official term given to the durational and spectral adjustments implemented when speaking in noise, was first reported in 1909 (Brumm & Zollinger, 2011). Various methods of speech rate modulations have been proposed and improved upon over the past 30 years to mimic this Lombard speech (Kupryjanow & Czyzewski, 2012a). Part of the reason for the lack of non-uniform algorithms may be that publication numbers on the Lombard effect have only started to increase in the late 2000's, resulting in a very gradual shift through relevant fields of study to eventually reach speech signal processing (Brumm & Zollinger, 2011). Nonetheless, there are very few papers proposing an automated method of non-uniform speech rate modulation.

However, there is a dire need for non-uniform speech rate modulation algorithms. The alternative to obtaining naturally sounding slow speech is to either record talkers who were speaking slowly (Iwasaki et al., 2002; Payton et al., 1994), or manually identify points where speech should be slowed down from its waveforms and spectrograms, and segmentally apply uniform speech rate modulation (Gordon-Salant et al., 2007). This is time consuming and requires specialized linguistics knowledge. Many of the non-uniform speech rate modulation algorithms that have been published also require training, and thus knowledge of linguistics such as the locations or the classes of the vowels (Kupryjanow & Czyzewski, 2012a) or are difficult to replicate due to poor documentation. Other algorithms are simple to implement, but produces unnaturally time scaled speech with uniform speech rate modulation.

The lack of literature on non-uniform speech rate modulation algorithms may be further confounded by the lack of objective speech intelligibility metrics applicable to time scaled speech. As such, past papers proposing a new non-uniform algorithm have reported through subjective human scores. However, using human subjects may not be possible on a limited time, budget, or speech database, where stimuli cannot be repeatedly used to compare algorithms. As such, only a few studies have tested their non-uniform speech rate modulation algorithms using speech recognition scores to evaluate intelligibility in human subjects (Jayan, Pandey, & Lehana, 2008; Kupryjanow & Czyzewski, 2012b). The majority have opted to measure speech quality as perceived by people, thereby allowing the use of duplicate speech stimuli (Demol et al., 2004; Grofit & Lavner, 2008; Lee, Kim, & Kim, 1997; Sreenivasa Rao & Vuppala, 2013). These latter studies may ask the listeners to rate which speech signals were most preferred, but do not

necessarily reflect how well it could be understood. Therefore, there is also a need for objective speech intelligibility metrics that can be applied to rate modulated speech to allow for the preliminary testing of these algorithms.

This thesis aims to advance the understanding of speech rate and intelligibility through a literature search and simulations involving speech enhancement algorithms that act through speech rate modulation as well as algorithms that enhance speech while keeping constant duration.

1.3 Overall Research Aim and Individual Research Objectives

The overall aim of this thesis is to advance the understanding of the association between speech rate and intelligibility especially in the context of noise and noise vulnerable populations. The specific research objectives of this thesis are as follows:

1. Identify existing objective speech intelligibility metrics, and suggest approaches by which they can be modified for use with rate modulated speech
2. Identify existing speech rate modulation based enhancement algorithms
3. Validate newly proposed metrics against existing objective speech intelligibility metrics under speech enhancement techniques which do not involve rate modulation, such as noise reduction
4. Explore the potential of newly proposed metrics and modified existing metrics in determining the speech intelligibility of the existing and newly proposed speech rate modulation based speech enhancement algorithms

Research objectives 1 and 2 are fulfilled in the literature review, Chapter 2, which aims to elucidate the gaps in the current literature. Research objectives 3 and 4 will be covered through empirical research, using the newly proposed metrics and algorithms introduced in Chapter 3, and research methodology explained in Chapter 4. The corresponding results are presented in Chapters 5, 6 and 7 for noisy, noise reduced and rate modulated speech. Finally, Chapter 8 details how each of the research objectives have been fulfilled by the thesis through the literature review and empirical findings, before offering some suggestions for future research directions.

1.4 Thesis Outline

Chapter 1 has given the background to the problem faced by people with hearing impairments, or elderly listeners in understanding speech in noise. One type of speech

enhancement, naturally adopted by humans, is to modulate speech rate non-uniformly. While algorithms that uniformly slow down speech have been studied extensively, there is limited research on non-uniform approaches. Additionally, there are no objective speech intelligibility metrics specifically for measuring improvements caused by rate modulation. The research focus and value, especially the problems caused by ignoring this health issue, were discussed, followed by the overall research aim and individual research objectives of this thesis.

Chapter 2 is a literature review. The section describes the benefits of rate modulation on speech comprehension, existing methods of rate modulation, and covers the concepts of existing methods to objectively measure speech intelligibility. Two key gaps in literature are identified in this chapter. Namely, the fact that little well documented non-uniform speech rate modulation algorithms exist, and that no previous study has attempted to objectively measure intelligibility of rate modulated speech. The strong association between speech rate and speech intelligibility is elucidated, which will be the basis of the newly proposed Speech Rate Based Intelligibility Metrics (SRBIM) and SOLely Acoustic Non-Uniform (SOLANU) speech rate modulation algorithm explained in detail in Chapter 3.

Chapter 3 presents the theory of the newly proposed SRBIM and SOLANU, as well as modifications to existing speech intelligibility metrics which makes them capable of handling rate modulated speech. The rationale for using speech rate to measure and improve speech intelligibility are given, as well as the details on how to derive SRBIM and SOLANU.

Chapter 4 covers the research methods involved in the empirical research. An explanation of the research strategy adopted and the rationale for the choice are provided. Details of the components necessary for data collection are covered, including the generation of several speech stimuli that has been noise reduced or appropriately rate modulated. The final section outlines the framework used for data analysis which involved analysis through correlation and scatter plots, a mock experiment that used new and existing metrics to compare the performance of the noise reduction algorithms, and some subjective observations for rate modulated speech.

The results will be presented in Chapters 5, 6 and 7. The results obtained during the optimization of the segmentation process are presented in Chapter 5. The correlation analysis of the metrics with noisy, but otherwise unprocessed speech will also be described in this chapter. Chapter 6 presents on the findings from similar analysis involving speech that had been noise reduced. Chapter 6 also includes mock experiment results of the metric agreements where the

performances of the three noise reduction algorithms were compared. Correlation analysis of rate modulated speech is given in Chapter 7. Some analysis involving my personal subjective observations of the intelligibility of rate modulated speech is also included in this chapter.

The final chapter, Chapter 8, is a summary of the findings as it ties back to the research aim and objectives set out in this chapter. Some recommendations for future research are presented.

1.5 Statement of Contribution

I have contributed to research in the field of speech signal processing in the following ways.

1. I have proposed and documented modifications that can be implemented on existing objective speech intelligibility metrics that work in the frequency domain to allow them to be used in the extended application of evaluating the speech intelligibility of rate modulated speech
2. I have proposed a new way of applying an existing objective measure of speech rate, STM-MFCC rates, and its variation of STM-LMPSC rates, as an alternative for speech intelligibility metrics called Speech Rate Based Intelligibility Metrics (SRBIM) with the specific aim of quantifying the performance of speech rate modulation based enhancement
3. I have compared these modified and newly proposed metrics in their performance with clean and noisy speech processed through three types of noise reduction algorithms, all of which constitute as a type of speech enhancement not involving time scaled speech
4. I have extended the comparison of these objective metrics with clean and noisy speech after implementing existing uniform and non-uniform speech rate modulation algorithms
5. I have further proposed SOLANU (SOLely Acoustic Non-Uniform) speech rate modulation algorithm as a simple method of generating non-uniform speech, and analyzed this algorithm under the mentioned metrics
6. I have programmed the MATLAB code for SRBIM and SOLANU available for use in future research. For researchers who would like to access this code, please refer to the journal publications below, which are affiliated with this thesis

In addition, I have contributed to research through presentations and publications.

1. I have published an abstract quantifying speech quality by comparison of the signal characteristics in the time domain

   Tanaka, M., & Smith, M. (presenter) (2010, July). *Identifying Appropriate Dynamic Time Warping, DTW, Metrics to Evaluate the Quality of Vocoder Time-Stretched Speech Phrases*. Poster presented at the 40th International Conference of the IEEE Engineering in Medicine and Biology Society, Honolulu, Hawaii.

2. I am currently preparing two articles based on this thesis. The titles of the articles are still tentative, and are subject to change

   Tanaka, M., & Smith, M. (2020). Speech Rate Based Intelligibility Metrics as a New Measure in Measuring Speech Intelligibility. Journal manuscript in preparation.

   Tanaka, M., & Smith, M. (2020). Derivation of the Stretch Function in Non-Uniform Speech Rate Modulation Using Spectral Transition Measures of Speech Parameters. Journal manuscript in preparation.

3. I have co-authored three technical transfer articles on the difficulties encountered in signal processing published in Circuit Cellar, a commercial magazine

   Smith, M., Farahani, E. S. & Tanaka, M. (2018, June). Murphy's Laws in the DSP World (Part 1): First Six Laws Explained. *Circuit Cellar*, 12-24.

   Smith, M., Tanaka, M., & Farahani, E. S. (2018, August). Murphy's Laws in the DSP World (Part 2): The Next Three Laws. *Circuit Cellar*, 14-27.

   Smith, M., & Tanaka, M. (2018, September). Murphy's Laws in the DSP World (Part 3): Future Imperfect. *Circuit Cellar*, 12-27.

2 Literature Review

Chapter 1 mentioned that people speaking in noise naturally slow speech down in a non-uniform manner to improve its intelligibility. This raises the question whether automatically generating slow speech will similarly improve intelligibility. While the literature contains many studies on methods of slowing speech down uniformly with respect to time, there are few speech enhancement algorithms that implement the more natural, human approach of slowing speech, which involves non-uniformly stretching the time scale. In addition, no previous studies have objectively measured intelligibility of such rate modulated speech.

This chapter will explore the support in literature in using speech rate modulation to improve speech intelligibility, the existing techniques for modulating the speech rate uniformly and non-uniformly, and the existing metrics for measuring speech intelligibility. The later sections will identify the problem with applying existing speech intelligibility metrics with rate modulated speech. This chapter aims to contribute to the current understanding of speech rate in association with speech intelligibility. Two gaps are identified in the literature pertaining to the lack of well documented non-uniform speech rate modulation algorithms, and the lack of objective means of measuring intelligibility of speech that has been processed through a rate modulation algorithm.

2.1 Effects of Speech Rate Modulation on Noisy Speech Intelligibility

In general, the literature supports the idea that slowed speech can overcome the negative effect of noise for listeners who are elderly or have hearing impairments. In the 1990's two studies showed that higher speech reception thresholds were observed with faster speech in noise in comparison to slower speech in noise for elderly listeners (Stollman & Kapteyn, 1994), and hearing or language impaired children (Stollman, Kapteyn, & Sleeswijk, 1994). This meant that speech in noise was more difficult to comprehend and needed to be presented at a higher signal to noise ratio (SNR) for faster speech for these noise vulnerable subjects. While a 37 % expansion relative to the original rate of speech was shown to not reduce the speech reception thresholds in children (Stollman et al., 1994), the study did not take into account the absolute rate of speech, making it difficult to rule out ceiling effects in comprehension. It is quite possible that the original speech rate was slow enough that maximum intelligibility in noise was achieved. If so, an increase in intelligibility could not be anticipated with a reduction in speech rate.

However, if the original speech had been recorded at a faster rate, it is possible that reducing its rate would have improved its intelligibility in noise. This speculation is supported in a more recent study implemented under three different speaking rates and two types of noise, where cochlear implant users exhibited the best level of comprehension with slow speech regardless of noise type (Shi et al., 2018). These studies support the notion that slow speech improves intelligibility for noise vulnerable listeners, even in the presence of noise.

While many studies are supportive of the benefits of a slow speech rate on intelligibility, some studies have shown contrary results. One study altered normal speech characteristics to mimic Lombard speech, a speech style adopted when speaking in a noisy environment (Cooke, Mayo, & Villegas, 2014). They found that increasing the duration of normal speech did not improve keyword recognition in the presence of stationary noise in normal hearing individuals, regardless of the variations of speech time scaling implemented (Cooke et al., 2014). Another similar study simulated cochlear damage by preprocessing speech through filters, and found that speech rate slowing did not improve speech intelligibility in noise (Nejime & Moore, 1998).

A consistent theme in the studies that did not observe benefits in intelligibility with slowing noisy speech down, was that the listeners were normal hearing people. This was most likely due to the difficulty of recruiting people who had hearing impairments for the experiments. In Chapter 1, the normal hearing listener's innate and robust ability to comprehend noisy speech, the cocktail party phenomenon, was introduced. Likewise, elderly listeners and hearing impaired listeners are populations known to have difficulties with understanding speech in noise. As a possible mechanism, the loss in temporal or spectral resolution of the brain in utilizing the transient releases from noise was mentioned. Thus, it is inappropriate to investigate how reducing the rate of speech for noisy speech does or does not improve comprehension levels if normal hearing test subjects are used. This is because their brains possess high auditory processing capabilities, and maximal comprehension was reached at the original speech rate presented, or perhaps at even faster rates. This hypothesis stands even when the speech stimuli is degraded to mimic hearing impairments (Nejime & Moore, 1998), as the difference in comprehension may be caused by problems in the brain, more so than in the ear for seniors.

Although there are some conflicting results regarding the benefits of speech rate modulation in improving speech intelligibility, many of the studies indicate that slowing speech

down makes it more easily understandable for people with hearing impairments or elderly listeners. The next section will describe the existing techniques used to modulate speech rate.

2.2 Method of Modulating Speech Rate

Two steps form the basis of speech rate modulation: analysis followed by synthesis of speech. Throughout these two procedures, the algorithm is expected to produce a high quality time stretched reconstruction of the original speech, whereby its pitch characteristics are preserved and sudden jumps in the phase of the output signal are avoided. For example, the sinusoidal framework models the speech as a sum of sine waves, and attempts to reconstruct a uniformly time scaled version by stretching out the individually identified sine waves by the desired stretch factor (Quatieri & McAulay, 1992). Time warping in conjunction with the Harmonic plus Noise Model (HNM) can also uniformly time stretch speech, albeit with some rather complex computations (Jayan et al., 2008). Phase vocoders are also capable of uniform speech rate modulation, by analyzing the speech in terms of frequency and phase (Götzen, Bernardini, & Arfib, 2000).

The most popular approaches used today for speech rate modulation are variants of the basic Overlap and Add, OLA, algorithm. This algorithm uses a fixed window size during analysis and synthesis which distorts the slowed speech by introducing an element of musical, or repetitive, noise. As the name implies, OLA algorithms analyze the speech signal in overlapping windows, which are sometimes duplicated and shifted before being added together in the time domain when synthesizing the output signal (Verhelst & Roelands, 1993). The shifting of windowed signals before combination produces a mismatch between the phases of the windowed signal, and the speech synthesized thus far, resulting in a beating noise with a period equal to the step size of the window. Synchronous Overlap and Add (SOLA) algorithms aim to reduce this introduced noise by aligning the signals at maximum points of correlation between the two combined signals (Turan & Erzin, 2015). Pitch-Synchronous Overlap and Add (PSOLA) does this by analyzing the pitch characteristics, especially in the voiced segments of the signal, and overlaps and adds signals by aligning these pitch markers (Moulines & Charpentier, 1990; Moulines & Laroche, 1995; Valbret, Moulines, & Tubach, 1992).

In contrast, Waveform Similarity Overlap and Add (WSOLA) ensures maximum alignment in the output signal by following, as closely as possible, the characteristics of the original signal as measured by cross-correlation coefficients, normalized cross-correlation

coefficients, or cross-AMDF (Average Magnitude Difference Function) coefficients (Demol et al., 2004; Grofit & Lavner, 2008; Verhelst & Roelands, 1993). As the succession of the development of these speech rate modulation algorithms suggest, PSOLA and WSOLA produce higher quality time stretched signals compared to their predecessors. Additionally, OLA type algorithms are flexible in their degree of stretching throughout the speech, enabling uniform speech rate modulation via a constant stretch factor, or non-uniform speech rate modulation by using a time varying stretch function.

The basic steps involved in uniformly rate modulating the speech are shown in Figure 2-1. This speech rate modulation algorithm assesses the pitch characteristics via an epoch extraction algorithm before overlapping and adding the signals in the time domain with a PSOLA algorithm (Murty & Yegnanarayana, 2008; Rudresh, Vasisht, Vijayan, & Seelamantula, 2018). As such, some of the papers call this algorithm Epoch Synchronous Overlap and Add (ESOLA) (Rudresh et al., 2018; Sreenivasa Rao & Vuppala, 2013).

2.2.1 Epoch Extraction

Epoch extraction is carried out by the zero frequency filtering method (Murty & Yegnanarayana, 2008), as shown in Figure 2-2. This method uses a specific type of filter called a resonator, which detects the impulse like signal excitation which occurs during the closing of the glottis, also known as the epoch (Murty & Yegnanarayana, 2008; Sao & Yegnanarayana, 2012; Yegnanarayana, Prasanna, & Guruprasad, 2011) . Since epochs are impulses, the energy of an

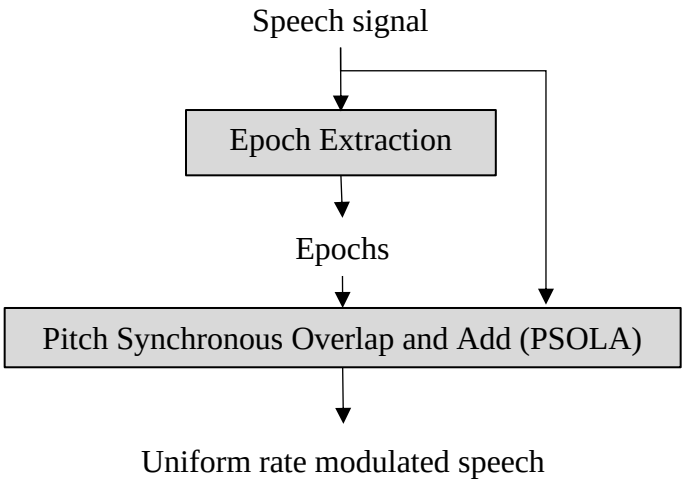


Figure 2-1: : Schematic for deriving uniform rate modulated speech using Pitch Synchronous Overlap and Add (PSOLA) algorithm. The two basic steps involved are the initial extraction of the epochs of speech, followed by the overlap and addition of the windowed segments of speech using the epoch locations for epoch alignment. The grey boxes contain multiple steps elaborated in later sections

epoch is spread out across all frequencies including the zero frequency (Sao & Yegnanarayana, 2012; Yegnanarayana et al., 2011). In comparison, the voiced speech has energy that is heavily concentrated on the frequency range unique to the voice, with only a small fraction located around zero frequency (Sao & Yegnanarayana, 2012; Yegnanarayana et al., 2011). Thus, implementing a zero frequency resonator to extract the epochs is advantageous as the vocal characteristics do not affect the epoch responses around zero frequency (Ni, Shiga, & Kawai, 2016). During epoch extraction, the speech signal is passed twice through the zero frequency resonator to further reduce the effects of the high frequencies (Murty & Yegnanarayana, 2008).

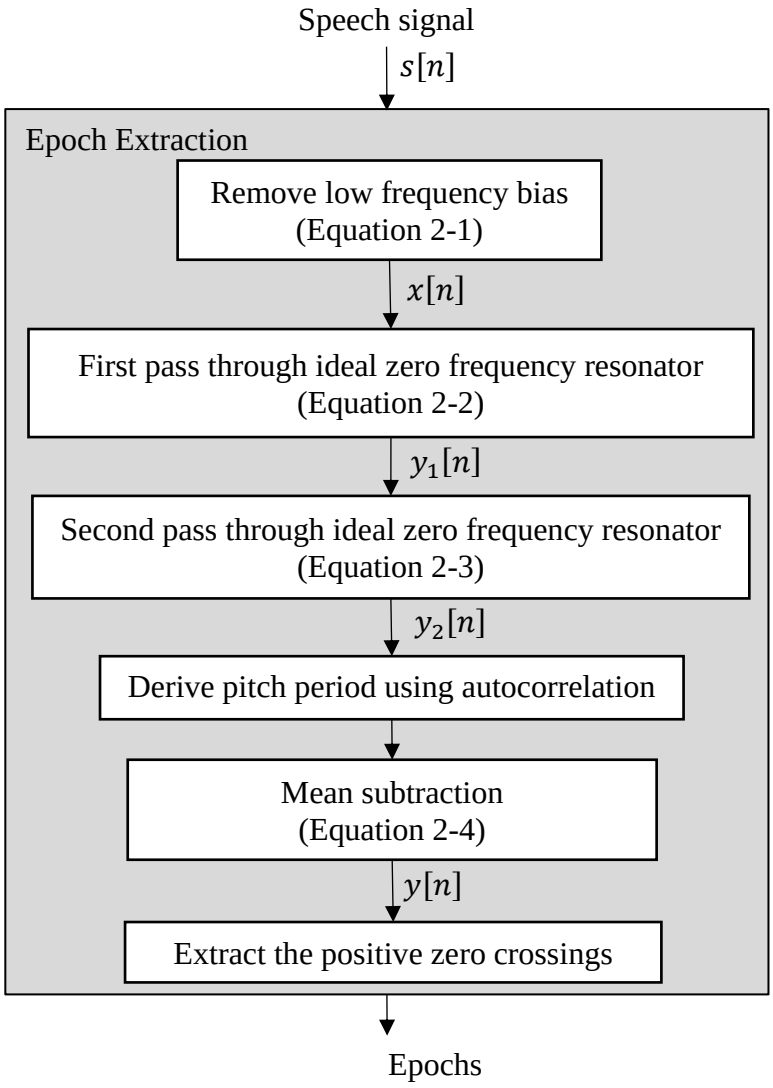


Figure 2-2: Schematic of extracting the epoch locations from the speech signal as the first part of speech rate modulation algorithm

The aim is to identify the epoch location, the instant of vocal tract excitation, which can be used as pitch markers (Murty & Yegnanarayana, 2008). This is done by first removing the difference in the speech signal,

$$x[n] = s[n] - s[n-1] \quad \textbf{(2-1)}$$

where $x[n]$ is the intermediate speech signal at sample point n, where the difference has been removed from the original speech $s[n]$ relative to the previous sampling point (Kadiri & Yegnanarayana, 2015; Murty & Yegnanarayana, 2008). The intermediate signal, $x[n]$, is then passed twice through the zero frequency resonator (Kadiri & Yegnanarayana, 2015; Murty & Yegnanarayana, 2008).

$$y_1[n] = -\sum_{k=1}^{2} a_k y_1[n-k] + x[n] \quad \textbf{(2-2)}$$

$$y_2[n] = -\sum_{k=1}^{2} a_k y_2[n-k] + y_1[n] \quad \textbf{(2-3)}$$

In the equations above, $a_1 = -2$ and $a_2 = 1$, and $y_1[n]$ and $y_2[n]$ are subsequent intermediate signals after passing through the filter once and twice, respectively (Kadiri & Yegnanarayana, 2015; Murty & Yegnanarayana, 2008). Then an autocorrelation function is applied to a 30 ms window of the speech to find the pitch period over which the local mean is removed (Kadiri & Yegnanarayana, 2015). The equation below shows the removal of the mean.

$$y[n] = y_2[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_2[n+m] \quad \textbf{(2-4)}$$

In this equation, $y[n]$ is the final signal where locations of its positive zero crossings, or the locations where the signal converts from a negative value to a positive value, are identified as epoch locations (Kadiri & Yegnanarayana, 2015; Murty & Yegnanarayana, 2008).

Once the epochs are identified, the next step in speech rate modulation is the overlap and add operation, implemented by the Pitch Synchronous Overlap and Add (PSOLA) algorithm.

2.2.2 Pitch Synchronous Overlap and Add (PSOLA)

The PSOLA algorithm is responsible for creating the rate modulated speech by overlapping and adding windows of the signals in the time domain. The steps involved in the

algorithm can be visualized in Figure 2-3. Epoch locations are used to determine the magnitude of the shift required during the analysis phase of the PSOLA algorithm. The actual shifting takes place during the synthesis stage.

First, the speech signal, $s[n]$, is split up into a series of overlapping analysis frames:

$$x_m[n] = s[n + mS_a], \qquad n \in [0, N-1] \tag{2-5}$$

where x_m is the rudimentary mth frame of analysis whose length is N, as indicated by the local index n within the window (Rudresh et al., 2018). This analysis signal is taken from a location in time corresponding to the frame index multiplied by the shift size of the analysis window, S_a

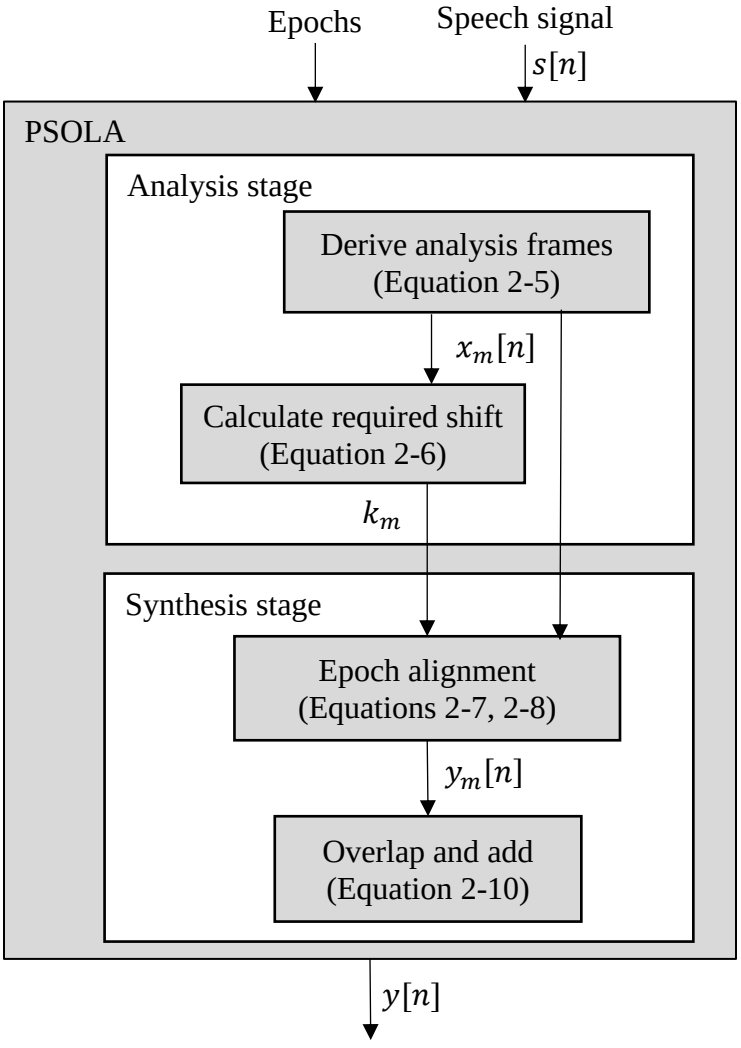


Figure 2-3: Schematic of Pitch Synchronous Overlap and Add (PSOLA) component that makes up the second part of speech rate modulation algorithm

(Rudresh et al., 2018). The analysis frames are rudimentary as they will be modified before the final synthesis. However, the initial analysis frames are necessary for determining the shift value, k_m, used in epoch alignment. Specifically, the region of overlap between the previous analysis frame and the current analysis frame is noted for the location and alignment of epochs. Keeping the previous analysis frame as is, the current analysis frame is designated a shift of k_m to ensure that the first epoch in the previous frame within this region of overlap coincides with the next upcoming epoch in the current analysis frame (Rudresh et al., 2018). Mathematically, this can be described as:

$$k_m = \min_{1 \le i \le P} \left(n_m^i - l_{m-1}^1 \right), \qquad where\ n_m^i > l_{m-1}^1 \tag{2-6}$$

where l_{m-1}^1 is the location index of the first epoch in the $(m-1)$th, or previous, analysis frame in the region of overlap $[(m-1)S_a + S_s, (m-1)S_a + N - 1]$, starting at a local index within the individual frame equal to the synthesis step size, S_s, and ending with the last point in the window, $N-1$. Similarily, n_m^i is the location index of the epochs in the mth analysis frame. The total number of epochs in this frame is denoted as P. The requirement for the epoch location index in the mth analysis frame is that it must be the closest epoch that occurs after the epoch index identified with l_{m-1}^1 in the region of overlap. If one or both of the preliminary versions of the analysis frames do not have an identified epoch, the mth synthesis frame, y_m, is equivalent to the mth analysis frame, x_m, (Rudresh et al., 2018).

$$y_m[n] = x_m[n], \qquad n \in [0, N-1] \tag{2-7}$$

Otherwise, the synthesis frame, y_m, is equal to the analysis frame, x_m, shifted by k_m (Rudresh et al., 2018).

$$y_m[n] = x_m[n + k_m] = s[n + mS_a + k_m], \qquad n \in [0, N-1] \tag{2-8}$$

During the synthesis phase, these frames of synthesis undergo overlap and add operations to yield the final rate modulated speech signal. The step size of the synthesis windows is:

$$S_s = \alpha S_a \tag{2-9}$$

where S_s is the synthesis step size, S_a is the analysis step size, and α is the stretch factor (Rudresh et al., 2018). This α value can be a constant stretch factor for uniform speech rate modulation, but can also be varied from frame-to-frame resulting in a stretch function in non-

uniform speech rate modulation. The synthesis step size also dictates the region of overlap when determining the shift of k_m. The overlap and add operations are implemented with the aid of a linear decay function β (Rudresh et al., 2018):

$$y[n + mS_s] = \begin{cases} \beta[n]y[n + mS_s] + (1 - \beta[n])y_m[n], & 0 \leq n \leq L - 1, \\ y_m[n], & L \leq n < N \end{cases} \quad \textbf{(2-10)}$$

where $n \in [0, L - 1]$ is the index within the region of overlap between the partially synthesized output signal, y, and the current synthesis frame begin overlaid and added, y_m. The length of the area of overlap, L, is simply the difference between the total window length, N, and the synthesis step size, S_s (Rudresh et al., 2018):

$$L = N - S_s \quad \textbf{(2-11)}$$

In producing the results presented in this thesis, the variant of Time Domain PSOLA (TD-PSOLA) (Moulines & Laroche, 1995) was used for modulating speech rate. TD-PSOLA is well documented, reliable and has been used in a previous non-uniform speech rate modulation algorithm (Sreenivasa Rao & Yegnanarayana, 2006). Its advantages include high quality output, appropriateness for time scale modification, low computational complexity in comparison to its Frequency Domain (FD) counterpart and consistency in comparing the uniform and non-uniform approaches used in this thesis. FD-PSOLA applies appropriate modifications in the frequency domain before transforming back to the time domain for the overlap and add operations (Moulines & Charpentier, 1990). As such, it is more computationally complex, and is more suitable for pitch modifications, rather than time scale modifications.

The TD-PSOLA variant explained in this section was used as a tool for rate modulating the speech, both in uniform and non-uniform approaches. The next section will describe existing techniques in non-uniformly rate modulating speech

2.3 Non-Uniform Speech Rate Modulation

Depending on the type of approach taken for modulating speech rate, non-uniform speech rate modulation may be achieved simply by incorporating a time varying stretch function rather than a constant stretch factor. In that regard, a new non-uniform speech rate modulation algorithm is sometimes synonymous with a novel method of deriving a stretch function. In past studies, different rationales and theories have been incorporated in obtaining the stretch function.

A simple approach to derive a stretch function is to vary the stretch according to the acoustic characteristics of speech. For example, expansion can be chosen to occur in only the voiced segments of the speech (Quatieri & McAulay, 1992). Alternatively, the speech signal may be analyzed for regions where the spectral characteristics change quickly, and the rate may be slowed down only for these transition segments (Jayan et al., 2008), or conversely for the remaining regions between the transition points (Grofit & Lavner, 2008; Lee et al., 1997). The advantage of acoustic approaches is in its simplicity and wide range of applicability. The algorithms would theoretically work as intended with any language, and the model is relatively simple to replicate.

An alternative approach for calculating a stretch function is to utilize the linguistic knowledge contained in the speech signal. Specifically, humans tend to stretch out vowels more than consonants when speaking slowly (Sreenivasa Rao & Vuppala, 2013). As such, complex algorithms may be designed to detect voice activity, then detect vowels and appropriately rate modulate the speech at the locations of the vowels (Kupryjanow & Czyzewski, 2012a). Alternatively, the algorithm may classify speech segments into vowels, pauses or consonants before further identifying the specific vowel type and vary the speech rate according to the vowel type (Sreenivasa Rao & Vuppala, 2013). These algorithms are more difficult to design compared to the acoustic stretch derivations as they require either knowledge of the language in the speech signal, such as vowel location and type, or require an abundance of training data that is labeled with similar information for developing an automatic vowel detector or classifier. Additionally, such algorithms will have limited applicability in a different language or with musical sounds and may be difficult to replicate without access to the same training data.

One existing algorithm attempts to incorporate the advantages of both acoustic and linguistic models. As shown in Figure 2-4, this approach aims to identify the linguistic characteristics of the speech through acoustic analysis (Demol et al., 2004). The algorithm first applies Hanning windows to the speech signal, and calculates the root mean square value of the speech signal, E_n (Kapilow, Stylianou, & Schroeter, 1999).

$$E_n = \sqrt{\frac{1}{N}\sum_{m=0}^{N-1} x^2[n+m]} \quad \textbf{(2-12)}$$

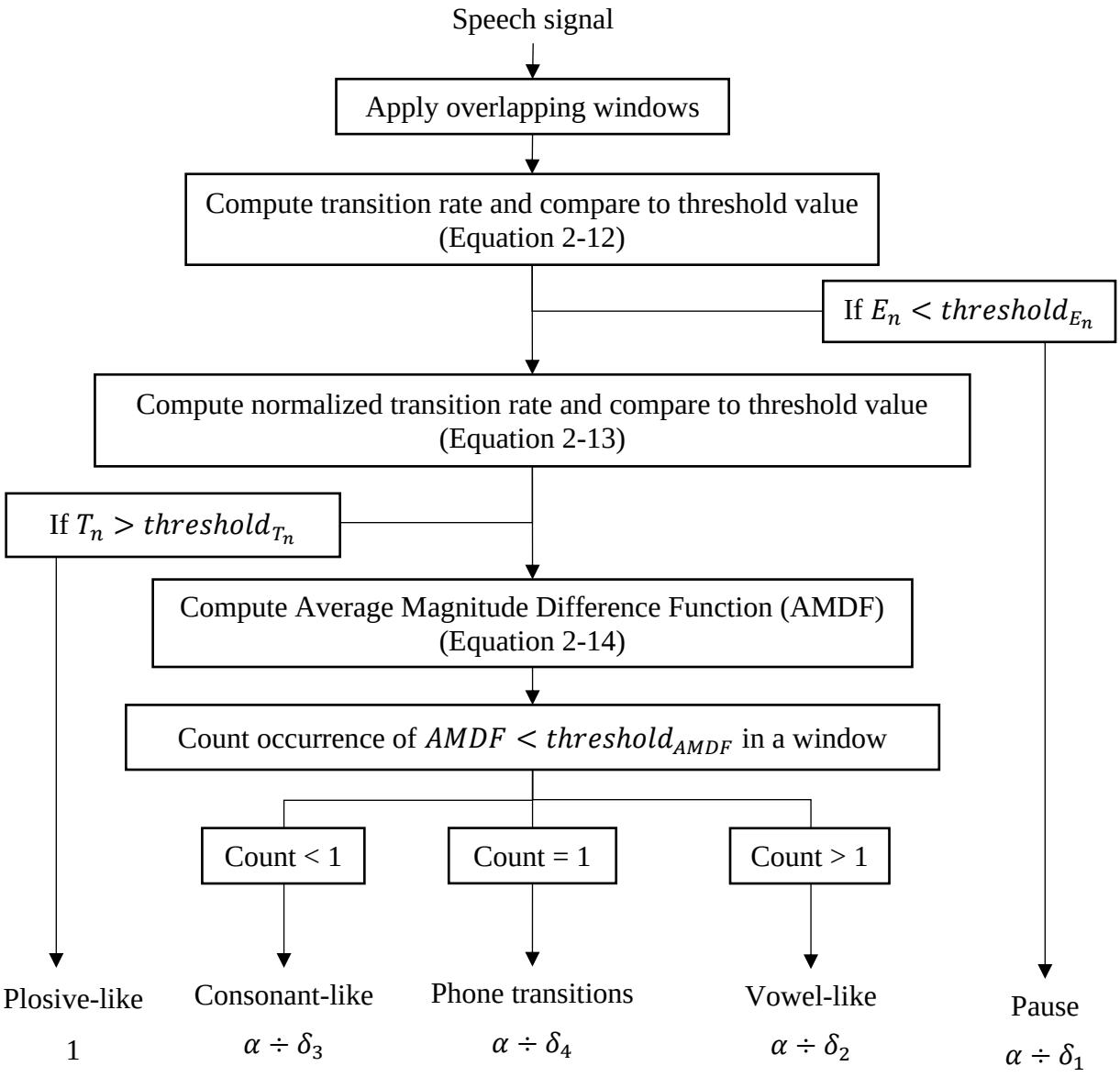


Figure 2-4: Flowchart of determining linguistic speech characteristics through acoustic analysis. The stretch value for each segment type is indicated as $\alpha \div \delta_x$, where $\delta_1 < \delta_2 < \delta_3 < \delta_4 < 1$ [Adapted from (Demol et al., 2004)]

In the equation above, $x^2[n+m]$ is the squared speech signal over one window as represented by the incrementing index of $m \in [0, N-1]$, and n corresponds to the start index of the window as designated by a user defined step size of the windows (Kapilow et al., 1999). If this root mean square value of the speech signal is below a user defined threshold, the algorithm classifies the current segment as a pause, and assigns it a stretch value of $\alpha \div \delta_1$, where α is the maximum stretch value used in any segment, and δ_1 is a value between 0 and 1 (Demol et al., 2004).

Next, the normalized transition rate of the root mean square value is calculated by:

$$T_n = \left| \frac{E_n - E_{n-1}}{E_n + E_{n-1}} \right| \quad \textbf{(2-13)}$$

where T_n is the normalized value, and E_{n-1} represents the root mean square value of the speech in the previous frame (Demol et al., 2004). If this normalized transition rate of the root mean square value of the speech signal is above another user defined threshold, the speech segment is labeled as plosive-like and given a stretch value of 1 (Demol et al., 2004).

The remaining segment types of vowel-like, consonant-like or phone transitions are classified by the periodicity in the Average Magnitude Difference Function (AMDF), taken over a longer window twice the length of the Hanning windowed signals (Demol et al., 2004).

$$AMDF(j) = \sum_{i=1}^{N} |x[i] - x[j+i]|, \qquad j \in [0, 2N-1] \quad \textbf{(2-14)}$$

In the equation above, the speech signal, x, is Hanning windowed over a length of N, but AMDF is taken over a longer window of length $2N$ (Demol et al., 2004). The speech segment is classified as vowel-like, consonant-like, or phone transition depending on how many times the AMDF dips below a third user defined threshold value (Demol et al., 2004).

In this manner, this non-uniform speech rate modulation algorithm classifies speech segments via acoustic analysis, then assigns corresponding stretch factors where α, a user defined maximum stretch factor, is adjusted by a δ_x value, such that $\delta_1 < \delta_2 < \delta_3 < \delta_4 < 1$ (Demol et al., 2004). This approach is simple to implement, but its performance is heavily dependent on the three user defined threshold values. If the thresholds are not appropriately scaled to the given speech signal, the speech segment classifier performance can be unbalanced. One shortcoming of this model is that the method of deriving these threshold values was not given (Demol et al., 2004).

One non-uniform speech rate modulation algorithm that is different from either acoustic or linguistic based approaches described above, is one that uses the glimpsing theory. This approach theorizes that people understand speech through the dips in fluctuating noise, and alters the speech rate non-uniformly to align the speech cues with these dips in noise (Cooke & Aubanel, 2017). The disadvantage of this method is that this requires both speech and noise signals to be known in isolation from each other. It is a rather difficult approach to incorporate

when the noise and speech are collected as one signal, as the sounds need to be separated before using this approach. As such, apart from noting that non-uniform speech rate modulation algorithms other than the two popular approaches do exist, it will not be mentioned further in this thesis.

A major shortcoming of existing models of non-uniform speech rate modulation is the difficulty of replicating the algorithms. One reason is the lack of documentation of the crucial details required. With the non-uniform speech rate modulation algorithm that mixed acoustic and linguistic models, the method of optimizing the user defined thresholds was never elucidated (Demol et al., 2004). With linguistic type non-uniform speech rate modulation algorithms, neither the detailed method of training the automatic vowel detector or classifier, nor the type of training data required, was documented (Kupryjanow & Czyzewski, 2012a; Sreenivasa Rao & Vuppala, 2013). This thesis will provide as much detail as possible to allow subsequent research on non-uniform speech rate modulation to follow without much unneeded challenges.

Once speech rate is modulated, its intelligibility needs to be measured in order to analyze whether a technique was successful in improving intelligibility. The next section will describe the existing metrics to objectively measure speech intelligibility.

2.4 Objective Speech Intelligibility Metrics

The ideal test of intelligibility of a noisy or processed speech signal would be to institute listening tests using the population of interest. However, these tests can be time consuming, costly and difficult to repeatedly recruit test subjects with the appropriate hearing status. An extensive speech database is usually required as it is unsuitable to use the same speech stimuli repeatedly with humans. Previous studies on rate modulated speech have also used subjective tests of asking listeners which time scaled speech was most preferred (Demol et al., 2004; Grofit & Lavner, 2008; Lee et al., 1997; Sreenivasa Rao & Vuppala, 2013), or quantified speech intelligibility through recognition scores (Jayan et al., 2008; Kupryjanow & Czyzewski, 2012b). Especially during the initial developmental stages of a speech enhancement algorithm, where many minor adjustments in design are anticipated, human testing is not realistic as an extensive number of speech stimuli and a great amount of time will be required to repetitively cycle through algorithm improvement and feedback. Instead, objective speech intelligibility metrics act as an inexpensive alternative to subjective listening tests.

Bell Telephone Laboratories were the first to attempt quantifying speech intelligibility through objective measures. In telecommunication experiments, it was found that different frequencies contributed to speech intelligibility differently (Fletcher, 1920). This observation was modeled into the Articulation Index (AI), computed as the weighted average of the signal to noise ratios (SNR) of designated frequency bands of the noisy speech (French & Steinberg, 1947). The American National Standards Institute later revised and improved AI as the Speech Intelligibility Index (SII) by changing the computations to manage distortions, masking and broadening of frequency bands (Kollmeier, Brand, & Meyer, 2008).

AI and SII require the long term spectra of the noise (Jensen & Taal, 2016). In other words, the metrics require the noise spectra to be known in isolation, which is often not available or becomes unavailable after the noisy speech has been enhanced through nonlinear modifications (Jensen & Taal, 2016). This is because nonlinearly modified noisy speech would contain noise that is also nonlinearly processed. Unlike additive noise, which is relatively simple to separate from speech through subtraction, nonlinearly modified noise would be extremely difficult to isolate. Additionally, AI and SII also rely on long term statistics, which means that two noises that are different in their presentation pattern with respect to time can be calculated to have equal intelligibility if their long term spectrum are identical, even though subjective tests reveal that intelligibility of these two noise types to be different (Jensen & Taal, 2016).

Since the introduction of the first objective metrics over 70 years ago, various modifications have been proposed and tested in the attempt to quantify the complex phenomenon of speech comprehension. It is difficult for one metric to represent speech intelligibility in all imaginable conditions and many metrics only focus on a specific situation or improve upon a certain limitation exhibited by a predecessor. For example, the speech transmission index (STI) was proposed as a metric that could account for some of the shortcomings observed with AI and SII, including nonlinear distortions such as peak clipping and reverberation (Steeneken & Houtgast, 1980). Subsequent modifications in STI enabled its use with nonlinear speech enhancement algorithms (Goldsworthy & Greenberg, 2004). More recently, speech-based envelope power spectrum model (sEPSM) generalized STI for speech with additive noise and reverberation (Jørgensen & Dau, 2011). Additionally sEPSM was also capable of measuring intelligibility of speech that had been enhanced by spectral subtraction or nonlinear noise reduction (Jørgensen & Dau, 2011). Similarly, SII was extended with Coherence Speech

Intelligibility Index (CSII) where the metric was made applicable to hearing aids and listeners with hearing impairments (Kates & Arehart, 2005).

The advantage of using traditional speech intelligibility measures lies in the large amount of empirical support that has accumulated over the years (Kollmeier et al., 2008). However, in more recent years, researchers have ventured into metrics derived from original lines of theory. For example, there is a speech intelligibility metric based on a psychoacoustic model, which uses correlation in the spectro-temporal parameters, or characteristics computed through a series of short term Fourier transforms (Christiansen, Pedersen, & Dau, 2010). Normalized Subband Envelope Correlation (nsec) also measured speech intelligibility as the similarity of the time-varying spectral features between the two speech signals (Boldt & Ellis, 2009).

In this thesis, simulation studies were performed using three existing speech intelligibility metrics proposed in the past ten years. The next sections will describe the method of deriving these measures.

2.4.1 Signal to Noise Ratio (SNR) Loss metric

Signal to noise ratio (SNR) loss was proposed as a speech intelligibility metric with the motivation to differentiate between spectral attenuation and amplification, two types of distortions that speech enhancement algorithms could potentially introduce (J. Ma & Loizou, 2011). As the name implies, SNR loss reflects the degree to which SNR has deteriorated compared to the clean speech. A value of 0 indicates perfect preservation of SNR characteristics, and the loss of intelligibility is capped at a value of 1 (J. Ma & Loizou, 2011).

To calculate SNR loss, the clean signal and the noisy or processed speech signals are required. Figure 2-5 shows a schematic of the steps involved in deriving SNR loss. The speech signals are first divided into smaller segments to be analyzed. This is done by applying a set of overlapping windows to the signal. For SNR loss, a Hamming type window was applied, and the short-time Fourier transform was calculated (Ma & Loizou, 2011. This frequency based value was then filtered through overlapping Gaussian-shaped filters, which represent the critical-band spectra observed in the ear (J. Ma & Loizou, 2011). The power within each band is summed up to derive a temporal-frequency representation of the speech signals.

$$Y(j,m) = X(j,m) + D(j,m) \quad \textbf{(2-15)}$$

In the equation above, j represents the frequency index corresponding to the ear's critical bands,

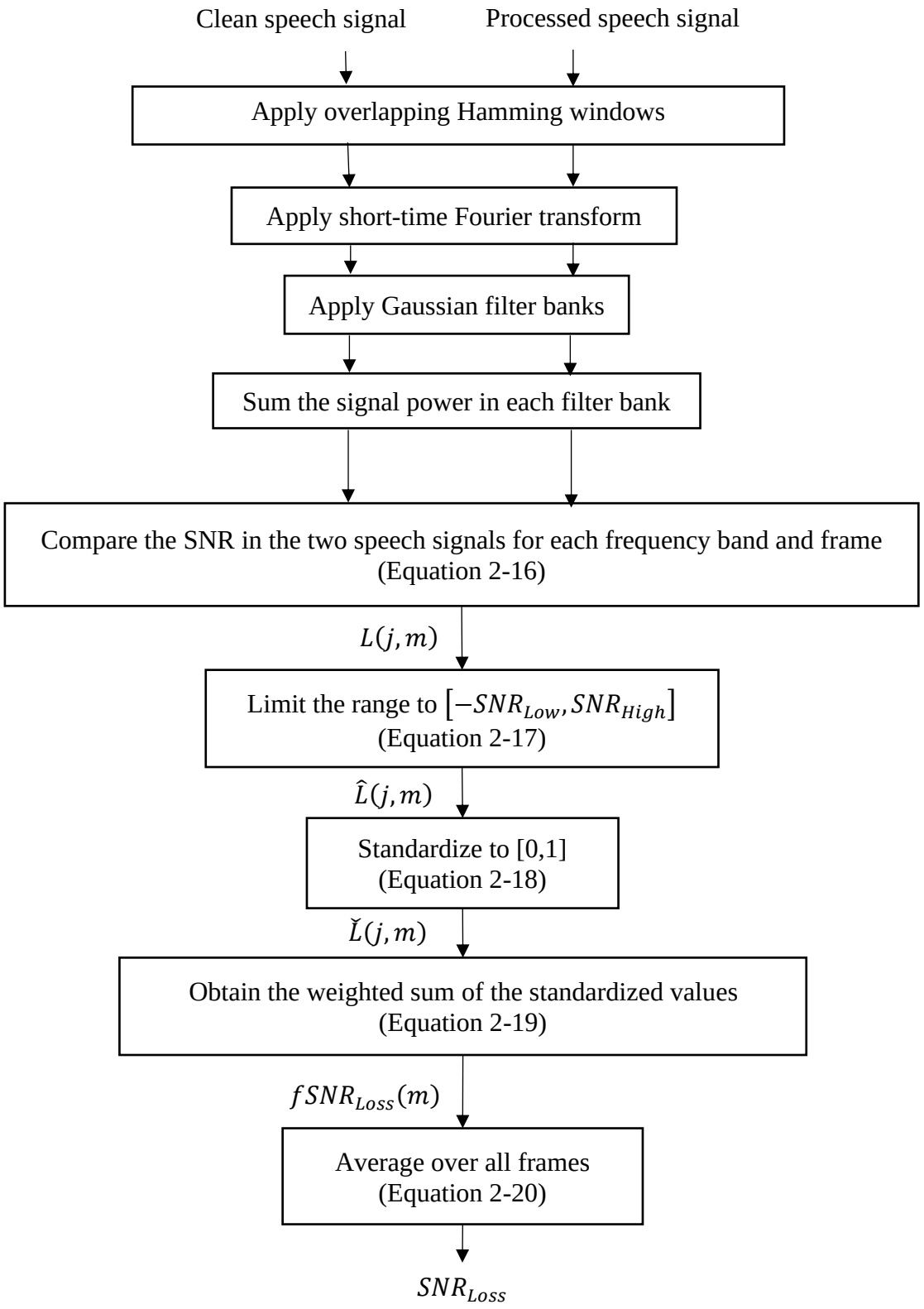


Figure 2-5: Schematic of the steps involved in calculating SNR loss intelligibility metric from the clean and processed speech signals

m is the frame index of the speech signals, and Y, X, and D are the spectra of the noisy speech, clean speech, and noise signal, respectively (J. Ma & Loizou, 2011). In this thesis, the term window is used when referring to its type, such as Hamming or Hanning windows, as well as when referring to one segment of speech where such a window has been applied. Frames, on the other hand, refer to the succession of windowed signals. If the windows were applied to the speech signal in an overlapping fashion, the resulting frames will also be overlapping. The temporal-frequency representation of speech is used to calculate the local SNR loss in each frequency band and frame.

$$\begin{aligned} L(j,m) &= SNR_X(j,m) - SNR_{\hat{X}}(j,m) \\ &= 10 \cdot log_{10} \frac{X(j,m)^2}{D(j,m)^2} - 10 \cdot log_{10} \frac{\hat{X}(j,m)^2}{D(j,m)^2} \\ &= 10 \cdot log_{10} \frac{X(j,m)^2}{\hat{X}(j,m)^2} \end{aligned} \quad \textbf{(2-16)}$$

In this equation, $\hat{X}(j,m)$ is the excitation spectrum of the processed speech, which can be visualized as the approximation of the spectrum of the clean speech, $X(j,m)$ (J. Ma & Loizou, 2011). The local SNR loss, $L(j,m)$, can be calculated by subtracting the SNR in frequency band j at frame m of the enhanced speech, $SNR_{\hat{X}}$, from that of the clean speech, SNR_X (J. Ma & Loizou, 2011). The SNR loss is further limited in its range from $[-SNR_{Low}, SNR_{High}]$ by:

$$\hat{L}(j,m) = \min\left(\max(L(j,m), -SNR_{Low}), SNR_{High}\right) \quad \textbf{(2-17)}$$

where $\hat{L}(j,m)$ is the range limited local SNR loss value (J. Ma & Loizou, 2011). The SNR loss is range limited based on the assumption that SNR of the noisy or processed speech is contained in this range, allowing for the precision of the measure to be maximized. For example, for noise reduced speech, a reduced SNR range of [-3, 3] dB outperformed the larger range of [-30, 10] dB when compared to subjective intelligibility via correlation (J. Ma & Loizou, 2011). The range limited local SNR loss values are further standardized into the range of [0, 1] using:

$$\check{L}(j,m) = \begin{cases} -\dfrac{C_-}{SNR_{Low}} \hat{L}(j,m), & if\ \hat{L}(j,m) < 0 \\ \dfrac{C_+}{SNR_{High}} \hat{L}(j,m), & if\ \hat{L}(j,m) \geq 0 \end{cases} \quad \textbf{(2-18)}$$

where C_- and C_+ are constants within the range of [0, 1] that control the degree to which amplification type and attenuation type distortions affect the final SNR loss metric, respectively (J. Ma & Loizou, 2011). The local SNR loss will then be weighted according to frequency band, to derive the frame SNR loss:

$$fSNR_{Loss}(m) = \frac{\sum_{j=1}^{K} W(j) \cdot \check{L}(j,m)}{\sum_{j=1}^{K} W(j)} \quad \textbf{(2-19)}$$

where K is the maximum frequency band index, $W(j)$ is the weight placed on frequency index j, and the frame SNR loss is represented by $fSNR_{Loss}$ (J. Ma & Loizou, 2011). The weights place importance on the different frequency bands, similar to the aforementioned frequency based contributions to speech intelligibility in AI and SII. This frame SNR loss is averaged to obtain a final SNR loss metric for the given speech segment:

$$SNR_{Loss} = \frac{1}{M} \sum_{m=0}^{M-1} fSNR_{Loss}(m) \quad \textbf{(2-20)}$$

where frame index m reaches the maximum frame count at $M - 1$, and SNR_{Loss} is the final SNR loss metric (J. Ma & Loizou, 2011).

2.4.2 Short-Time Objective Intelligibility (STOI) Measure

Traditional objective speech intelligibility metrics have the advantage of having empirical support in its implementation. At the same time, many of these traditional measures have a large number of parameters determined through experimentation, leading to a loss in transparency (Taal, Hendriks, Heusdens, & Jensen, 2010). The motivation for the development of short-time objective intelligibility (STOI) measure was to propose a simple objective intelligibility measure, while addressing the lack of agreement with subjective ratings with certain speech enhancement algorithms, such as ideal time frequency segregation (Taal et al., 2010; Taal, Hendriks, Heusdens, & Jensen, 2011).

As shown in Figure 2-6, to derive STOI, first, the clean and processed speech are resampled at 10 kHz, applied overlapping Hanning windows, and converted into the frequency domain through a Fourier transform (Taal et al., 2010, 2011). If the energy contained in a frame is below 40 dB compared to the maximum energy frame in the clean speech, the frame is considered to be silent, and removed from further analysis (Taal et al., 2011). Then, the

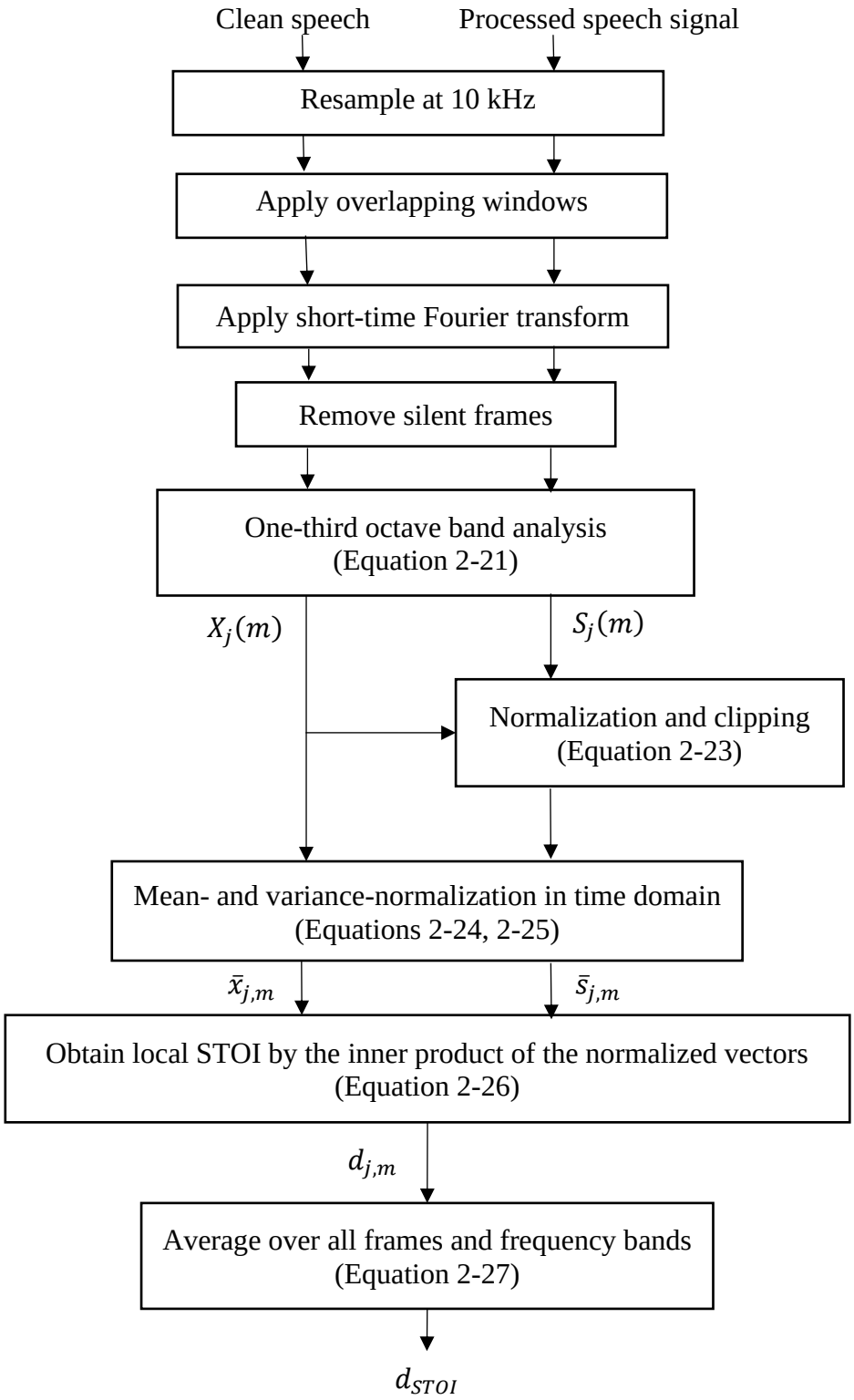


Figure 2-6: Schematic of the steps involved in calculating the STOI metric from the clean and processed speech signals

frequency bins are analyzed under one-third octave bands, and its norm, labeled as the time-frequency unit, is calculated (Taal et al., 2010, 2011).

$$X_j(m) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} |\hat{x}(k,m)|^2} \quad \textbf{(2-21)}$$

To clarify, norm refers to a term used in linear algebra, defining the length of a multidimensional vector (Nicholson, 2006). In Equation 2-21, $\hat{x}(k,m)$ represents the kth frequency index in frame m of the speech signal, which are taken from jth one-third octave band limits of $[k_1(j), k_2(j) - 1]$ (Taal et al., 2010, 2011). The processed speech signal's time frequency unit was calculated in a similar manner as the clean speech, and is represented as $Y_j(m)$ (Taal et al., 2010, 2011). To calculate the correlation of the clean and processed speech over time, the short-time temporal envelope, $x_{j,m}$, can be represented in vector format:

$$x_{j,m} = \begin{bmatrix} X_j(m-N+1) \\ X_j(m-N+2) \\ \vdots \\ X_j(m) \end{bmatrix} \quad \textbf{(2-22)}$$

where N is the length of the analysis window (Taal et al., 2011). The equivalent short-time temporal envelope of the processed speech ($s_{j,m}$) requires normalization and clipping ($\hat{s}_{j,m}$), to ensure unintelligible segments are upper bound limited to β = -15 dB (Taal et al., 2011).

$$\hat{s}_{j,m}(n) = \min\left(\frac{\|x_{j,m}\|}{\|s_{j,m}\|} s_{j,m}(n), \left(1 + 10^{-\frac{\beta}{20}}\right) x_{j,m}(n)\right) \quad \textbf{(2-23)}$$

In this equation, n is the local index number in the range of $[0, N-1]$ (Taal et al., 2011). Both clean and processed speech signals are then mean- and variance-normalized in each frequency band index j with:

$$\bar{x}_{j,m} = \frac{1}{\left\|x_{j,m} - \mu_{x_{j,m}}\underline{1}\right\|}\left(x_{j,m} - \mu_{x_{j,m}}\underline{1}\right) \quad \textbf{(2-24)}$$

where $\underline{1}$ is an vector of all ones, and $\mu_{x_{j,m}}$ is the sample mean calculated by (Jensen & Taal, 2016):

$$\mu_{x_{j,m}} = \frac{1}{N}\sum_{n=0}^{N-1} X_j(m-n) \quad \textbf{(2-25)}$$

The processed speech, after normalization and clipping ($\hat{s}_{j,m}$), is also mean- and variance-normalized. The local STOI ($d_{j,m}$) can then be calculated as the inner product of the normalized vectors:

$$d_{j,m} = \bar{s}_{j,m}^{T}\bar{x}_{j,m} \quad \textbf{(2-26)}$$

where $\bar{s}_{j,m}^{T}$ is the transpose of the mean- and variance-normalized vector of the processed speech $\bar{y}_{j,m}$ (Jensen & Taal, 2016). To derive the final STOI value (d_{STOI}), the average of this intermediate metric is required over all frequency bands and frames:

$$d_{STOI} = \frac{1}{JM}\sum_{j.m} d_{j,m} \quad \textbf{(2-27)}$$

where J and M are the total number of one-third octave bands and frame numbers involved in the speech signal analysis (Taal et al., 2010, 2011).

2.4.3 Extended Short-Time Objective Intelligibility (ESTOI) Measure

As the name of the extended STOI (ESTOI) measure suggests, this metric modified STOI to better handle fluctuating noise (Jensen & Taal, 2016). Noise that masks the desired speech is often called noise maskers as it covers and distorts the target speech. This is achieved by computing spectral correlation coefficients, and averaging this over time within analysis windows, rather than taking the inner products of the normalized temporal envelopes which was done with STOI (Jensen & Taal, 2016). Figure 2-7 gives the schematic of the steps involved in calculating ESTOI. The additional normalization procedure allows ESTOI to quantify intelligibility of speech, even in fluctuating noise.

The steps involved in calculating ESTOI are similar to those involved in calculating STOI. The clean and processed or noisy speech signals are resampled to 10 kHz, applied a series of overlapping Hanning windows, converted to the frequency domain, removed of the silent frames, and applied the one-third octave band analysis (Jensen & Taal, 2016). For clarity of the subsequent steps, the spectral vectors can be represented as a matrix:

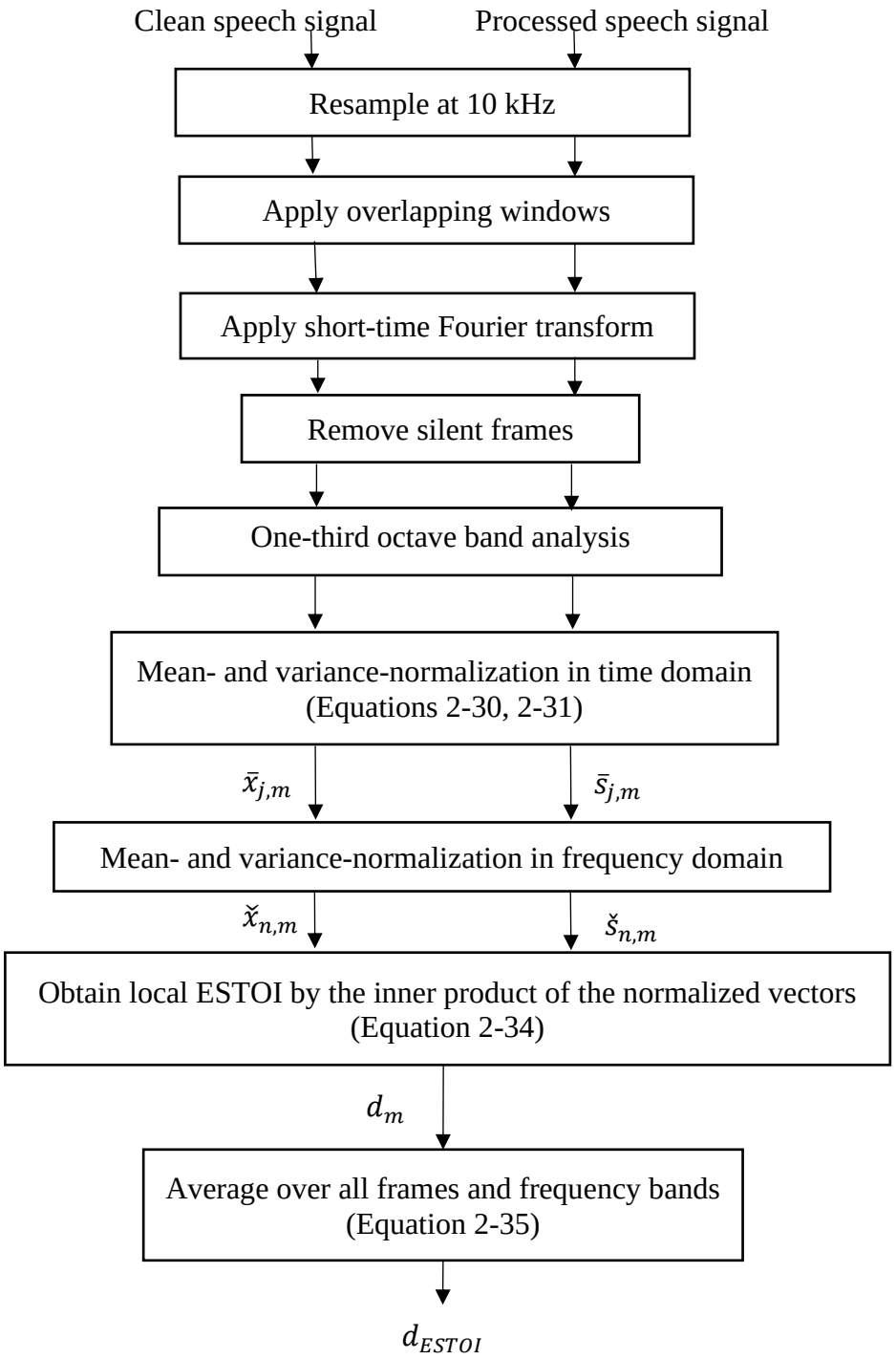


Figure 2-7: Schematic of the steps involved in calculating ESTOI metric from the clean and processed speech signals

$$S_m = \begin{bmatrix} S_1(m-N+1) & \cdots & S_1(m) \\ \vdots & \cdots & \vdots \\ S_j(m-N+1) & \cdots & S_j(m) \\ \vdots & \cdots & \vdots \\ S_J(m-N+1) & \cdots & S_J(m) \end{bmatrix} \quad \textbf{(2-28)}$$

where the subscript j represents the one-third octave band index $[1, J]$ and m is the frame index within the whole speech segment, and the local frame index within each row of the matrix varies from $[0, N-1]$ corresponding to the analysis window (Jensen & Taal, 2016). As with STOI, each frequency index is mean and variance normalized, which is equivalent to normalizing each row in Equation 2-28. One row in S_m represents the temporal envelope of the speech signal in the jth frequency band:

$$s_{j,m} = [S_j(m-N+1) \quad S_j(m-N+2) \quad \cdots \quad S_j(m)] \quad \textbf{(2-29)}$$

which is normalized via equations similar to Equation 2-24 and Equation 2-25:

$$\bar{s}_{j,m} = \frac{1}{\left\| s_{j,m} - \mu_{s_{j,m}} \underline{1} \right\|} \left(s_{j,m} - \mu_{s_{j,m}} \underline{1} \right) \quad \textbf{(2-30)}$$

$$\mu_{s_{j,m}} = \frac{1}{N} \sum_{n=0}^{N-1} S_j(m-n) \quad \textbf{(2-31)}$$

This row-wise normalization can be envisioned as the temporal normalization of each one-third octave frequency band, and was also observed in STOI calculations. For ESTOI, spectral normalization is also implemented on the columns (Jensen & Taal, 2016). After row-normalization, the spectrogram matrix represented in Equation 2-28 can be described as:

$$\bar{S}_m = \begin{bmatrix} \bar{s}_{1,m} \\ \vdots \\ \bar{s}_{j,m} \\ \vdots \\ \bar{s}_{J,m} \end{bmatrix} \quad \textbf{(2-32)}$$

where each row ($\bar{s}_{j,m}$), starting from $\bar{s}_{1,m}$ and ending at $\bar{s}_{J,m}$, is derived from Equation 2-30 (Jensen & Taal, 2016). Column-normalization on this spectrogram matrix produces zero-mean normalized spectra, which was temporally normalized prior to normalization across the frequency indices, which can be represented as:

$$\check{S}_m = [\check{s}_{0,m} \quad \cdots \quad \check{s}_{n,m} \quad \cdots \quad \check{s}_{N-1,m}] \tag{2-33}$$

In this equation, $\check{s}_{n,m}$ represents the column-normalized spectrogram at the local frame index n within the range $[0, \mathrm{N}-1]$ for all frequency indices, which makes up the final row- and column-normalized spectrogram matrix, $\check{S}_m$ (Jensen & Taal, 2016). The intermediate intelligibility index at frame index m can then be calculated as:

$$d_m = \frac{1}{N}\sum_{n=0}^{N-1} \check{s}_{n,m}^T \check{x}_{n,m} \tag{2-34}$$

where $\check{s}_{n,m}$ and $\check{x}_{n,m}$ are unit vectors representing the spectrogram at local frame index n for the processed and clean speech, respectively (Jensen & Taal, 2016). A simple average of these local intelligibility measures over the frames results in the final ESTOI metric:

$$d_{ESTOI} = \frac{1}{M}\sum_{m=1}^{M} d_m \tag{2-35}$$

Column normalization produced unit vectors for the processed (or noisy) and clean speech signals. Multiplying these two vectors in Equation 2-34 can be visualized as the orthogonal projection of the target speech signal upon the reference speech signal (Jensen & Taal, 2016). A value of 1 would mean perfect alignment, while -1 would mean a complete reversal in directionality. As such, ESTOI has a maximum range of [-1, 1], with much of the measures taking on the range of [0, 1].

2.5 The Problem of Existing Speech Intelligibility Metrics

The previous section described several existing speech intelligibility metrics. However, no previous study has applied these metrics to the rate modulation type speech enhancement techniques, which is the focus of this thesis. Many existing metrics require perfect time alignment between the clean reference speech signal and the compared noisy or processed signal. However, after rate modulation, these signals are no longer of equal length and the standard metrics fail.

However, when considering existing speech intelligibility metrics that convert the speech signals into the frequency domain, it is theoretically possible to compare speech segments of different lengths, as long as they represent the same segment relative to the whole, and assuming

that the applied rate modulation was ideal and did not introduce changes to the frequency characteristics of the speech. The three existing speech intelligibility metrics described in detail in the previous section was chosen precisely because these measures compare the speech signals in the frequency domain. As such, simple modifications can be introduced to make it applicable to rate modulated speech, which will be covered in Chapter 3.

No previous study has applied existing objective speech intelligibility metrics to measure intelligibility of rate modulated speech. Similarly, no metric has been proposed with the specific aim of measuring speech intelligibility based on speech rate. The next section will provide the support in literature for using speech rate to quantify intelligibility.

2.6 Speech Rate and Its Associations with Speech Intelligibility

The previous section outlined existing speech intelligibility metrics, explaining three measures in detail. These three metrics compare the target and references speech in the frequency domain, and we have proposed that they could be altered for use with time scale modified speech. Additionally, in Chapter 3, this thesis will introduce new metrics for measuring time stretched speech based on speech rate. This section will provide the rationale and support for using speech rate to measure speech intelligibility.

Literature indicates that slower speech results in higher intelligibility while faster speech decreases intelligibility. Several studies support the idea that slower speech results in improved intelligibility for listeners, especially those with hearing impairments. Performance significantly improved with slower speech in a study testing ten cochlear implant users and their speech recognition as a function of speech rate, (Iwasaki et al., 2002). In this study, speech recognition scores for fast, medium and slow speech were 15.7 %, 39.0 % and 56.0 % ($p < 0.0001$), respectively, revealing that speech rate had a significant influence on speech perception in cochlear implant users (Iwasaki et al., 2002). In fact, slower speech rate may be associated with higher speech intelligibility in elderly or hearing impaired individuals more so than in normal hearing individuals. With normal hearing older adults, sentence perception improved when the speech rate was modulated to be 60 % slower (Cho et al., 2014). In a different study, while natural speech rate showed excellent speech recognition scores across all test subjects and time expansions, selective slowing of pre-compressed speech improved speech intelligibility for elderly listeners and those with hearing loss (Gordon-Salant et al., 2007). In the presence of noise, elderly listeners with hearing impairments exhibited better sentence recognition with

slower speech (Lessa & Costa, 2013). These studies indicate that slow speech is beneficial for listeners that are elderly or hearing impaired, the same populations that were also vulnerable to noise.

Conversely, increasing speech rate reduces speech comprehension, especially in the presence of noise. Even in young normal hearing listeners, speech intelligibility of Chinese Mandarin speeches declined rapidly from when stimuli was presented at 400 words per minute to 1200 words per minute (Du, Lin, & Wang, 2014). In one study, speeding up noisy speech decreased the speech identification score in older listeners (Gordon-Salant & Fitzgibbons, 2004). Interestingly, while listener's age was correlated with speech recognition of rate modulated noisy speech, hearing status was not a factor in this study (Gordon-Salant & Fitzgibbons, 2004). Additionally, for older listeners with hearing loss, noisy fast speech was especially susceptible to temporal envelope distortions introduced by a hearing aid (Jenstad & Souza, 2007), implying that the use of a hearing device has additional negative effects on intelligibility of noisy speech. Fast speech seems to result in lower comprehension in the same populations as those who benefitted with slower speech. This effect was evident in fast speech combined with noise.

From analyzing the literature of experiments that modulated speech rate and observed its effects on intelligibility, the same group of listeners that had difficulty understanding speech in noise also benefitted from the provision of slow speech and suffered in comprehending fast speech. This may be because fast speech and noisy speech encounter the same bottleneck in the auditory processing capabilities of these individuals. Namely, auditory processing is slower, and the appropriate speech cues cannot be picked out for comprehension. It is possible that normal hearing seniors have a reduced speed in processing sound, sampling the speech sparsely in comparison to younger listeners. Rate modulation to slow down speech may be able to compensate for their slower central auditory processes either through increased opportunities to pick up speech cues, or more time to process each cue. Likewise, hearing impaired listeners who exhibit a reduced spectro-temporal resolution in hearing may be able to better collect relevant speech cues from slowed down speech resulting in improved comprehension. Nonetheless, the literature indicates that decreasing speech rate was especially beneficial in terms of intelligibility for noise vulnerable listeners, indicating an association between rate and intelligibility.

This link between speech rate and intelligibility suggests that measuring speech rate may reflect intelligibility of a given speech signal. In fact, a study revealed an interesting correlation

between hearing performance in noise and hearing performance with faster speech. Similar speech recognition decreases were observed in older listeners who listened to speech with progressively greater compression ratios (or greater speech rate), as when they were provided with speech with incrementing amounts of noise (Kumar, Pradhan, Handa, & Sanju, 2016). Hence, with appropriate matching of speech rate to noise level, speech rate could act as a substitute measure for speech intelligibility in certain cases.

2.7 Pertinent Issues

This chapter identified the support for using speech rate modulation to improve intelligibility and described existing methods of uniformly and non-uniformly modulating speech rate, and presented existing objective speech intelligibility metrics of SNR loss, STOI and ESTOI, that could be modified for use with rate modulated speech. The details on how these frequency based metrics will be modified will be covered in Chapter 3. The intricate link between speech rate and intelligibility was then presented as rationale for using speech rate as the basis for a new intelligibility metric.

The literature review undertaken identified two pertinent issues in speech rate modulation. One issue was the need for new non-uniform speech rate modulation algorithms that were transparent in its implementation, while providing a natural sounding output, rather than a uniformly rate modulated speech. While humans naturally slow down speech by non-uniformly modulating the speech rate, the literature on speech rate modulation type enhancement algorithms, especially those on non-uniform approaches, is limited. There are more research on uniform approaches, but these algorithms yield speech output that is unnatural and sounds almost “drunken” (Quatieri & McAulay, 1992). However, many of the previous non-uniform algorithms are complex, and difficult to replicate, sometimes missing important information needed in its implementation.

The second issue was the need for objective speech intelligibility metrics for use in the preliminary developmental stages of rate modulation type enhancement algorithms. Previous studies in speech rate modulation have used human subjects to either rate their listening preferences of various time stretched speech (Demol et al., 2004; Grofit & Lavner, 2008; Lee et al., 1997; Sreenivasa Rao & Vuppala, 2013), or speech intelligibility through measures such as recognition scores (Jayan et al., 2008; Kupryjanow & Czyzewski, 2012b). As human testing is costly and time-consuming, a more convenient option is to use objective speech intelligibility

metrics. However, research on objective measures of speech intelligibility in the context of rate modulated speech does not exist, including studies on how to modify existing metrics to cater to rate modulated speech, and studies that propose new metrics that are specifically designed for applicability to rate modulated speech.

The next chapter will describe the theory of the newly proposed speech intelligibility metrics based on speech rate, and the new non-uniform speech rate modulation algorithm.

3 Theory of Proposed Algorithms

Intelligibility of the spoken words can be increased by slowing it down. Such speech rate modification can be achieved in a uniform manner, or in a more natural fashion through non-uniform rate modulation. The literature review illustrated two issues in the field of speech rate modulation type enhancement techniques. The first issue was the lack of objective intelligibility measures capable of quantifying speech intelligibility of rate modulated speech. The second issue was the general lack of non-uniform speech rate modulation algorithms, complicated by the difficulty to implement and use the available algorithms due of poor documentation.

In response, modifications are documented for existing speech intelligibility metrics to make them capable of handling rate modulated speech. In addition, two new Speech Rate Based Intelligibility Metrics (SRBIM), specifically designed to measure the intelligibility of rate modulated speech, are proposed. Furthermore, a new SOLely Acoustic Non-Uniform (SOLANU) speech rate modulation algorithm is described, capable of generating a non-uniformly rate modulated speech. This chapter documents the modifications implemented on existing metrics, as well as the step-by-step derivation of the new SRBIM and SOLANU algorithms.

3.1 Modifying Existing Speech Intelligibility Metrics

This section will briefly review the three existing speech intelligibility metrics used to investigate uniform rate modulated speech, and explain the modifications to be applied to extend their capability to handle rate modulated speech.

ESTOI (Extended Short-Time Objective Intelligibility), STOI (Short-Time Objective Intelligibility) and SNR (signal to noise ratio) loss are three existing speech intelligibility metrics that were introduced in detail in Chapter 2. They require the noisy or processed speech signal that is to be evaluated, and the clean version of the speech for reference. These measures have been found to correlate well to subjective scores from volunteers for noisy speech (Taal et al., 2010) and noise reduced speech (J. Ma & Loizou, 2011). In its original form, these metrics required complete time alignment between the target and reference speech. In other words, while they were capable of measuring intelligibility of speech that had been enhanced spectrally, the metrics could not measure intelligibility of speech that had undergone any modifications in the time domain. However, this thesis proposes a method to adapt these metrics to handle speech

signals that have been stretched in time, by taking advantage of the fact that they compare the target and reference speech in the frequency domain.

Rate modulated speech that has been slowed down for speech enhancement is longer than its original version. As such, metrics that quantify speech intelligibility in the time domain cannot handle rate modulated speech, as the number of sample points have changed relative to the original signal. However, metrics that compare in the frequency domain can be modified. The proposed approach is to appropriately elongate the window length and step size, making frequency based metrics capable of outputting an intelligibility measure for rate modulated speech.

ESTOI, STOI and SNR loss all measure intelligibility through spectral analysis. This process involves the conversion of segments of the speech into the frequency domain, which makes the metric flexible to some speech signal differences in time between the target and reference speech, as long as the frame of the target time stretched speech is representative of the corresponding frame in the clean speech. For example, consider a speech signal that has been time scaled so that it is twice the length of the original speech. By analyzing the elongated speech with a window that is double the length of the original analysis window, the speech segments should have comparable frequency characteristics as the reference speech, under the implied assumption that the pitch has been properly preserved in the rate modulation process. In analyzing these signals digitally, this step can be interpreted as taking a signal segment in time that is longer than the reference speech, but when converting it into the frequency domain, the resolution of the spectral representation of the elongated signal is adjusted to be equivalent to that of the reference spectrum. This is achieved by increasing the step size of the overlapping analysis windows for the elongated speech so that the total number of frames in the two speech signals become equivalent. This allows a one-to-one frame wise comparison of the two speech signals, allowing ESTOI, STOI and SNR loss to measure intelligibility of rate modulated speech.

In essence, appropriately adjusting the window length and the step size used in framing the speech can make these frequency based metrics capable of measuring intelligibility of rate modulated speech. However, this only applies well for uniformly rate modulated speech. With non-uniformly rate modulated speech, the elongated speech segments would no longer align with the reference speech segments. A severe underestimation of the actual intelligibility of the slow speech would be anticipated. In the next section, the thesis follows up on SRBIM, a possible

solution to this problem through a new intelligibility metric specifically designed to handle non-uniformly modified speech rates.

3.2 Logic Behind Speech Rate Based Intelligibility Metrics (SRBIM)

The literature review showed that hearing impaired individuals seem to be more sensitive to changes in speech rate, demonstrating reduced speech comprehension with faster speech. From this relationship between speech rate and intelligibility, this thesis proposes that speech rate has the potential to act as an intelligibility metric. Rate modulation algorithms aim to improve speech intelligibility, making speech rate an excellent metric in the initial assessment of whether that goal was achieved. In this section, a new method of measuring speech intelligibility, Speech Rate Based Intelligibility Metrics (SRBIM), will be described. Two variants of SRBIM, MFCC-SRBIM (Mel Frequency Cepstral Coefficient type Speech Rate Based Intelligibility Metric) and the LMPSC-SRBIM (Log Mel Power Spectral Coefficient type Speech Rate Based Intelligibility Metric), are detailed.

Much like many metrics exist for measuring speech intelligibility, many algorithms for objectively measuring speech rate have also been proposed. Starting with the pioneering speech rate metrics of enrate (Morgan, Fosler, & Mirghafori, 1997) and mrate (Morgan & Fosler-Lussier, 1998), there are many possible candidates for SRBIM. In this thesis, normalized STM (Spectral Transition Measure) rates based on boundaries detected from the changes of MFCC and LMPSC was used as SRBIM. One of the measures, MFCC-SRBIM, was based on the existing speech rate metric, STM-MFCC rate.

STM-MFCC rate calculates speech rate by counting the number of detected speech boundaries based on the rate of change in MFCC, and divides this by the duration of the speech signal. The STM of the cepstral coefficients of speech are able to identify locations of sudden change in frequency characteristics, which are presumed, in the literature, to be the boundaries in speech (Aharonson et al., 2017; Dusan & Rabiner, 2006; Furui, 1986). STM-MFCC rate can be calculated by dividing the number of segments, separated by these boundaries, over the total duration of the speech.

Since speech rate metrics are used to measure changes in intelligibility in this thesis, the change in the metrics in comparison to a reference signal was used, rather than the absolute rate value of an individual target speech. This is because the absolute speech rate value only reflects the speed, or how fast or slow, a speech signal is, rather than give an indication of how the noise

or the speech processing has affected it. In other words, the important information is whether the speech rate has changed as a result of the added noise or after processing it through a speech enhancement algorithm, in comparison to the speech rate measured for its clean reference speech. This new normalization step of taking the ratio between the absolute STM-MFCC rate values of the target and clean speech, is similar to how existing intelligibility metrics compare noisy speech to its reference speech. The final SRBIM measure resulting from this ratio was called MFCC-SRBIM.

MFCC-SRBIM was chosen for the ability of STM-MFCC to detect speech transitions that agree with expert opinion. The original paper on STM (Furui, 1986) documented how sudden changes in STM of cepstral coefficients, before the implementation of the new normalization step, have high alignment with perceptually critical points that greatly affected keyword recognition. When compared to phone, or speech boundaries, manually determined by experts in 4620 sentences, STM-MFCC showed correct detection 84.6 % of the time, missed expert identified boundaries 15.4 % of the time, and inserted extra boundaries 28.2 % of the time (Dusan & Rabiner, 2006). This rate is comparable to the agreement of manually annotated boundaries among experts. In a study measuring the agreement rate of manually identified speech boundaries in four analysts, consisting of two qualified speech therapists, one linguist and one psychologist, only 66.03 % of 471 boundaries were identified by all members (Oehmen, Kirsner, & Fay, 2010). The highest agreement was seen for 76.92 % of the boundaries which were successfully detected by three or more of the analysts (Oehmen et al., 2010). The 84.6 % agreement of the STM-MFCC based boundary detection with manual annotations is comparable to this agreement rate among experts. The high agreement rate of STM-MFCC in identifying speech boundaries implies that it can detect the frequency of the changes in speech patterns which are important for speech comprehension, making it an ideal SRBIM.

The second SRBIM used in this thesis, the LMPSC-SRBIM, is a slight variant of the above. LMPSC-SRBIM is calculated in the same manner as MFCC-SRBIM, while using an intermediate parameter, LMPSC, instead of MFCC. This intermediate set of parameters have been used in Automatic Speech Recognizers (ASR), especially with the aim to improve ASR robustness to additive or reverberant noise (Krueger & Haeb-Umbach, 2010; Leutnant, Krueger, & Haeb-Umbach, 2014; Li, Wang, Zhou, Bourlard, & Liao, 2013). The second SRBIM was chosen to reduce the problems associated with noise. One of the anticipated issues was that

SRBIM would be excessively reactive to noise, reaching a ceiling value for many of the noisy speech samples. This would make it useless in quantifying differences in speech intelligibility over a certain threshold. Thus, LMPSC-SRBIM was included in this thesis for its usefulness in making ASR robust to noise.

3.3 Deriving Speech Rate Based Intelligibility Metrics (SRBIM)

Calculation of the two proposed SRBIM consists of four basic steps. Figure 3-1 shows these steps starting with the clean reference signal and the target signal, which is either noisy or speech enhanced. In the following sections, the individual steps involved in calculating the two proposed SRBIM will be described in more detail. Section 3.3.1 covers the first step of deriving the appropriate parameters (first block in Figure 3-1). Section 3.3.2 describes the next block in Figure 3-1, the method of converting these coefficient values into its Spectral Transition Measure (STM). The next block in Figure 3-1 is explained in Section 3.3.3, where the spikes in the STM are detected through a peak picking algorithm. The last step of normalizing the rates, the last block in Figure 3-1, is depicted in Section 3.3.4, resulting in the desired modified SRBIM value.

3.3.1 Mel Frequency Cepstral Coefficient (MFCC) and Log Mel Power Spectral Coefficient (LMPSC) Derivation

The first step in calculating the two SRBIM is to obtain the appropriate speech parameters to be used for subsequent calculations. The two parameters are Mel Frequency Cepstral Coefficients (MFCC), a well-known parameter in speech signal processing, and Log Mel Power Spectral Coefficients (LMPSC), an intermediate set of values obtained while deriving MFCC.

Mel Frequency Cepstral Coefficient (MFCC) is a speech parameter that has been researched extensively. Conceptually, it quantifies the spectral contour of the speech by treating the signal in the frequency domain by operations typically applied in the time domain (De Leon & Martinez, 2012; Schafer, 2008). It was developed to effectively represent acoustic information in a concise manner, without losing crucial information required for speech recognition (S. B. Davis & Mermelstein, 1980). This was based on the assumption that speech varies slowly relative to the carrier frequency of the digital signal, and its characteristics would be maintained within a 20 ms to 30 ms window (Schafer, 2008). As such, a series of short-term time and frequency analyses on these windows would result in an accurate sketch of the speech (Schafer, 2008). MFCC is significant as it models cochlear perception of sound through the adoption of the

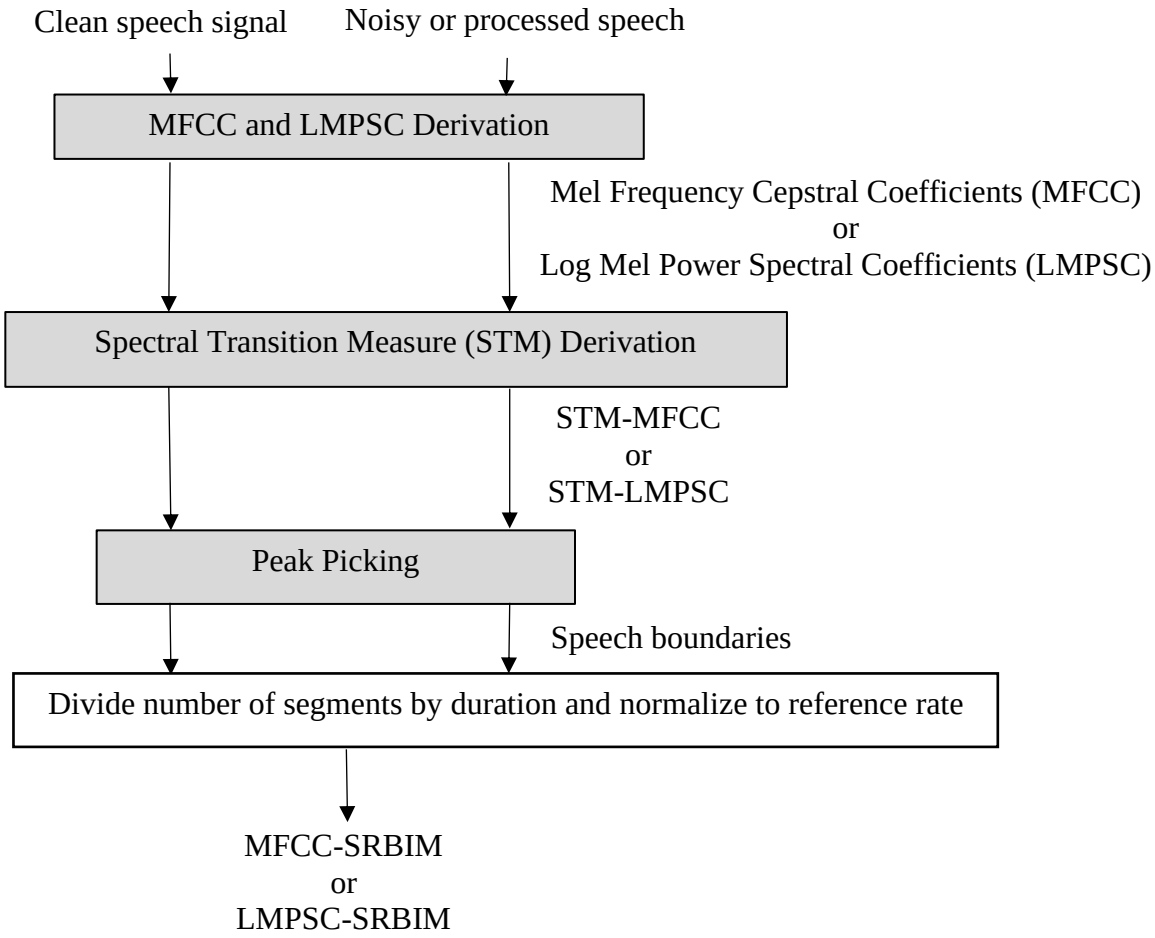


Figure 3-1: Schematic for deriving MFCC-SRBIM or LMPSC-SRBIM. Grey boxes consist of several steps covered in detail in later sections

mel scale, a filter bank with center frequencies and bandwidths based on human sensitivity (Schafer, 2008). The number and shape of filter banks, as well as the number of calculated cepstral coefficients, can vary between MFCC variants. Typically, anywhere between twenty and twenty-four triangular filters are used with 8 kHz speech signal, and ten to nineteen cepstral coefficients are derived, with thirteen and nineteen being common choices in studies (S. B. Davis & Mermelstein, 1980; Schafer, 2008; Zheng, Zhang, & Song, 2001). Today, MFCC is one of the commonly used representation of speech signals, especially when dealing with perception of the signal by a human observer (Schafer, 2008).

Different methods of calculating MFCC and its intermediate LMPSC exist, but the general steps are common between the variants. Figure 3-2 shows the steps involved in deriving

the two parameters. The first objective in deriving the coefficients is to obtain the short-time power spectrum of the reference and target speech. In order to calculate this, speech is first divided into a series of overlapping frames. A Hamming window is applied to each frame to smooth the signal, before converting it into the frequency domain using the short-time Fourier transform. This results in the short-time power spectrum of the signal.

The next aim is to analyze the signal contents in the frequency domain through spectral and cepstral coefficients. A set of triangular mel frequency filter banks are applied to each frame and used to generate log energy coefficients, or LMPSC in the frequency (spectral) domain. This can be further converted into the quefrency (cepstral) domain for MFCC through a discrete cosine transform, resulting in the two sets of speech parameters required for the next step in deriving SRBIM.

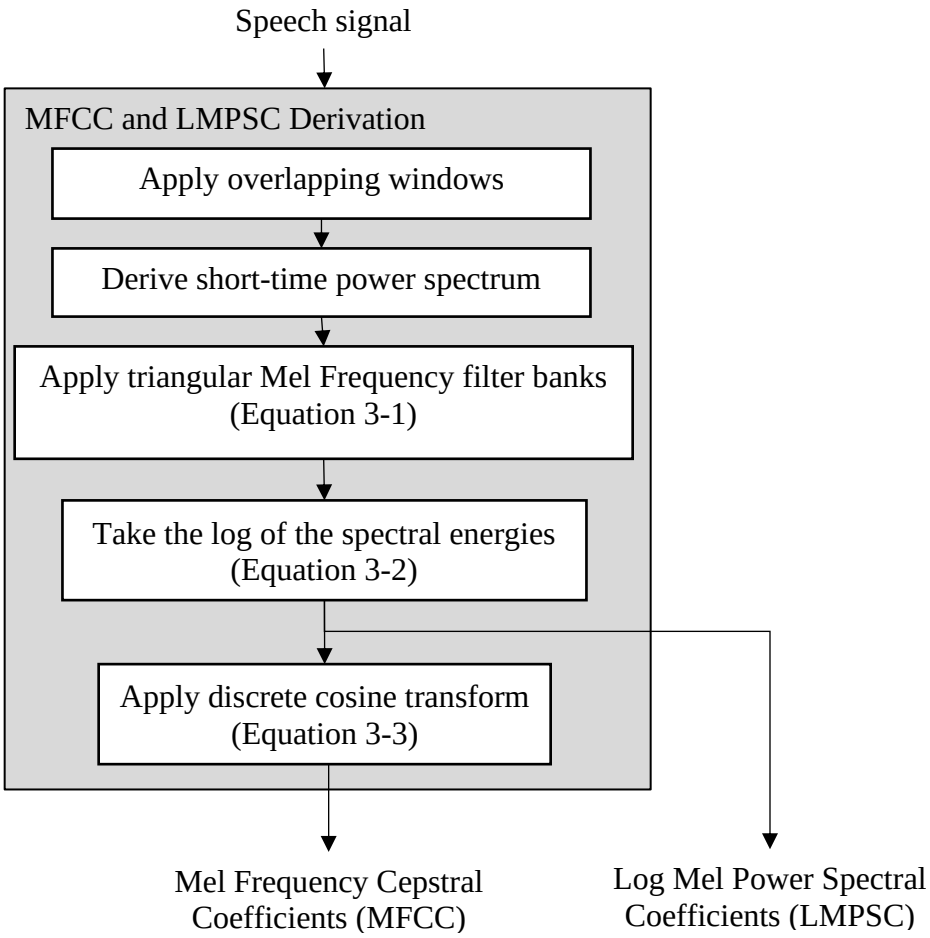


Figure 3-2: Schematic for deriving Mel Frequency Cepstral Coefficients (MFCC) and Log Mel Power Spectral Coefficients (LMPSC)

It has been mentioned that MFCC has several variants, much like the objective speech intelligibility metrics and speech rate measures covered in Chapter 2. In this thesis, the MFCC calculation that was documented in the Hidden Markov models (HMM) Toolkit (HTK) was used (Young et al., 2009). The differences between the various MFCC derivations lie in the details, such as how the mel frequency scale is defined, how many filters are used in the filter bank, whether a log or natural log is used in logarithmic compression, and how the decorrelation using discrete cosine transform is implemented. With the MFCC derivation used in this thesis, the mel scale was defined as:

$$\hat{f}_{mel} = 2595 \cdot log_{10}\left(1 + \frac{f_{lin}}{700}\right) \quad \textbf{(3-1)}$$

where $\hat{f}_{mel}$ was the warped frequency in Mels, and f_{lin} was the linear frequency in Hz (Ganchev, 2011). For logarithmically compressing the filter output, this model takes the natural logarithm:

$$LMPSC_i = ln\left(\sum_{k=0}^{N-1} |S(k)|^2 \cdot H_i(k)\right), \quad i = 1, 2, \dots, M \quad \textbf{(3-2)}$$

where $H_i(k)$ was the filter weight at frequency index k for the set of triangular filters, $S(k)$ was the spectral representation of the speech and $LMPSC_i$ was the spectral output for the ith triangular filter out of a total of M filters after log compression with the natural ln operation (Ganchev, 2011). The MFCC was obtained by decorrelating this output through a discrete cosine transform:

$$MFCC_r = \sqrt{\frac{2}{M}} \sum_{i=1}^{M} LMPSC_i \cdot cos\left(\frac{r(i - 0.5)\pi}{M}\right), \quad r = 0,1, \dots, R - 1 \quad \textbf{(3-3)}$$

to calculate the rth MFCC ($MFCC_r$) for a total of R coefficients (Ganchev, 2011). Once spectral and cepstral coefficients were calculated for the reference and target signals, the next step was deriving its Spectral Transition Measures (STM).

3.3.2 Spectral Transition Measure (STM) Calculation

After features were calculated for each frame, its Spectral Transition Measures (STM) were used to detect the boundary points within the continuous speech. STM is the rate of change in speech spectral characteristics, originally proposed as a method for detecting boundaries

between vowels and consonants (Furui, 1986). Sudden changes in spectral features characterize speech segmentation points, which shows as spikes in the STM (Dusan & Rabiner, 2006). Perceptual experiments indicated that accurate perception of a concatenated syllable was dependent on the location of where the syllable was cut, relative to the STM spike (Furui, 1986). The first half of the syllable was correctly identified by an observer when the cut was 2.6 ms or more after the maximum transition point, while the second half was recognized when the syllable started 4.0 ms or more before the maximum transition point (Furui, 1986). These results indicated that a small 10 ms window centered on the STM spike held crucial speech cues required for comprehension (Furui, 1986).

While there are various ways to quantify the rate of change in the spectral qualities of speech, STM in this thesis is based on the magnitude of the rate of spectral change. STM derivation based on the rate of spectral change can be divided into two steps, as shown in Figure 3-3. The regression coefficients of the relevant spectral or cepstral parameters are obtained, which are simplified into one dimension for the STM.

Mathematically, the regression coefficient was calculated by:

$$a_i(n) = \frac{\sum_{m=-I}^{I} v_i(n+m) \times m}{\sum_{m=-I}^{I} m^2} \quad \textbf{(3-4)}$$

where $a_i(n)$ was the regression coefficient of the i^{th} spectral feature over a range of frames

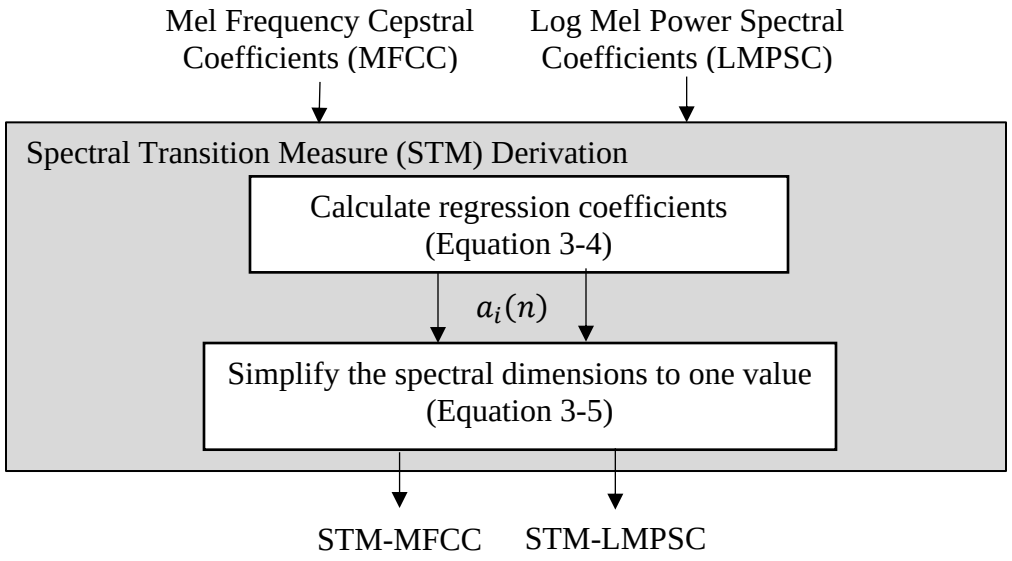


Figure 3-3: Schematic for deriving Spectral Transition Measure (STM) of speech parameters

centered around the current frame n, m was the number of frames on either side of the current frame under analysis, ranging from a value of $-I$ to I, and $v_i(n)$ was the i^{th} MFCC or LMPSC value (Dusan & Rabiner, 2006; Furui, 1986). Typically, I of 2 has been used in previous studies where frames were incremented by 10 ms, resulting in a 50 ms window over which the regression coefficient was calculated (Dusan & Rabiner, 2006).

The regression coefficient is a measure of the changes in the chosen spectral feature through a small collection of neighbouring frames. It is calculated for each dimension within the spectral feature, corresponding to the filter index for LMPSC, or the coefficient number for MFCC. STM is a method of collapsing these individual measures of spectral change for all D dimensions, resulting in one simple measure representing the magnitude of the rate of change in the spectral or cepstral feature as a function of time. This collapsing of the regression coefficients into one value can be represented mathematically as:

$$STM(n) = \frac{(\sum_{i=1}^{D} a_i^2(n))}{D} \quad \textbf{(3-5)}$$

where D was the number of spectral features, such as the number of MFCC or LMPSC used, and $STM(n)$ is the STM value at frame index n (Dusan & Rabiner, 2006; Furui, 1986). The derived STM for the reference and target speech are then processed through a peak picking algorithm to identify their speech boundaries.

3.3.3 Peak Picking

STM measure of spectral or cepstral coefficients represents the magnitude of the spectral rate of change. Spikes in the STM measure indicate sudden changes in the speech characteristics and are potential speech boundaries that can be identified through a peak picking algorithm. However, not all STM peaks are speech boundaries. Raw STM spikes are considered to be candidates for speech boundaries, and STM in conjunction with threshold algorithms determine the actual transition points, eliminating any spurious or false boundaries (Aharonson et al., 2017; Dusan & Rabiner, 2006). In this thesis, median based adaptive threshold peak picking algorithm was chosen because thresholding based on local medians are less affected by outliers, such as large peaks in the signal (Bello et al., 2005). Peak picking algorithms that are affected by large peak values tend to miss smaller peaks that occur near a large peak. For that reason, the median filter is known to have better adaptivity than the average filter (Kauppinen, 2002).

The general steps for finding the peaks in the STM are shown in Figure 3-4. There were two user defined parameters, the window length and the threshold factor, that were needed for the median based adaptive threshold peak picking algorithm to work. The window length dictated the range of sample points over which the median value was calculated from. The median was then multiplied by the threshold factor to yield the adaptive threshold. These parameters were empirically optimized through simulations, which will be covered in more detail in Section 4.2.5. STM values above the adaptive threshold value were identified as candidates for speech boundaries. The peak candidates were then analyzed in groups. Each group consisted of all consecutively identified speech boundary candidates, defined at both ends by a dip in the STM value below the adaptive median threshold. Then, the location where the largest difference between the adaptive threshold and the STM signal was identified as a peak, and subsequently, a boundary point in speech. All other peak candidates in the same group were labeled as non-peak.

The median based adaptive threshold peak picking algorithm identified peaks in both the target and reference signals, effectively segmenting the speech. The number of segments was converted into rate of speech and normalized, ultimately resulting in a normalized rate of speech, which was the SRBIM value.

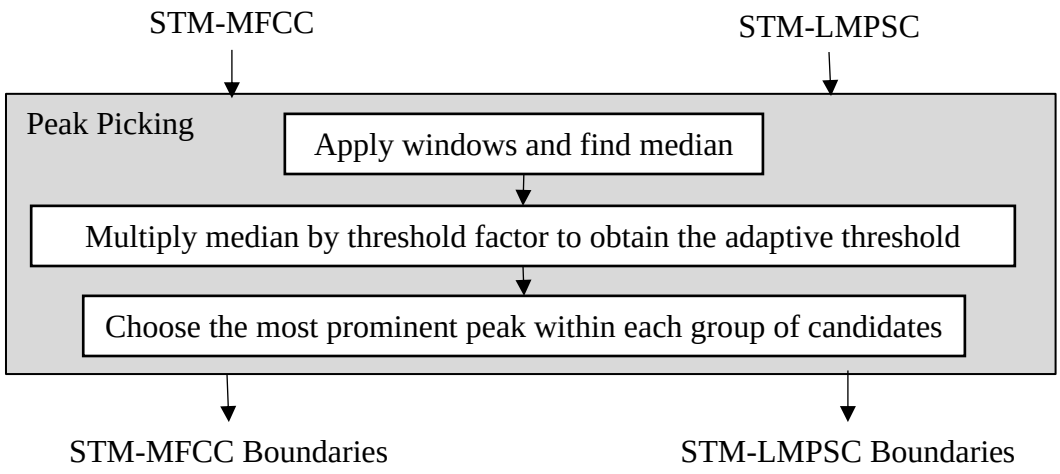


Figure 3-4: Schematic for detecting speech boundaries from STM with a peak picking algorithm

3.3.4 Rate Normalization

The peak picking algorithm identified the location of the transition points in speech, which was interpreted as the boundaries in speech. These speech boundary points were assumed to be separating segments in speech, which were tallied as number of speech units over a known duration. As such, the number of peaks was converted into the number of speech segments and divided by the total time of the speech signal to result in a speech rate measure.

The intent of SRBIM was to use speech rate as an intelligibility measure. As such, the STM-MFCC or STM-LMPSC speech rates calculated thus far do not have much meaning. Rather, these rates need to be normalized to a reference speech. Normalization is a new step proposed in this thesis in order to calculate the ratio between the analyzed speech and the clean speech. This value better represents the change in speech rate, and consequently, intelligibility. For example, consider two different speech signals in noise, where one is detected to have greater STM-MFCC rate than the other for the clean, noiseless condition. Even with the presence of additive noise, it is quite possible that the same relative outcome will be observed, with the same speech stimuli outputting a greater STM-MFCC rate than the other speech. One may be misled that the noisy speech that produced a greater STM-MFCC rate experienced a greater decrease in intelligibility as a result of the noise, which may not be true. However, by normalizing the STM-MFCC rate of both noisy speeches to the rate derived with their clean versions, a comparable normalized rate, which is the SRBIM, may be observed between the two. In summary, the rate normalization method standardizes the output, making it possible to compare SRBIM across different speech stimuli.

Section 3.3 has described the two newly proposed Speech Rate Based Intelligibility Metrics (SRBIM) of MFCC-SRBIM and LMPSC-SRBIM in detail. The next section will describe the new non-uniform speech rate modulation algorithm, which utilizes many of the steps involved in calculating SRBIM.

3.4 SOLely Acoustic Non-Uniform (SOLANU) Speech Rate Modulation

The new algorithm was named SOLANU (SOLely Acoustic Non-Uniform) speech rate modulation, to pay tribute to the PSOLA algorithm used to produce the rate modulated speech, while emphasizing the fact that the stretch function was derived through an acoustic approach to produce a non-uniformly modulated output. This algorithm was proposed because the literature on non-uniform speech rate modulation algorithms were limited, and those that were available

sometimes lacked crucial details, making it difficult to replicate the algorithms. Thus, there was a need for a new algorithm that balanced the desired elements of algorithm transparency and naturalness of the output speech. The acoustic model was chosen for the basis of the algorithm, for its simplicity and wide range of applicability. A recurring theme in existing acoustic approaches to non-uniform speech rate modulation was the concept of varying the stretch factor according to the locations of the transients of the speech (Grofit & Lavner, 2008; Jayan et al., 2008; Lee et al., 1997). As linguistic models commonly slowed the vowel segments while keeping the consonants relatively quick (Kupryjanow & Czyzewski, 2012a; Sreenivasa Rao & Vuppala, 2013), the desired speech output was postulated as a signal that preserved the timing information contained in the quickly changing transient segments of the speech, while slowing down the regions between these transition points. In that regard, the new non-uniform speech rate modulation algorithm was similar in approach to two of the previous acoustic based algorithms (Grofit & Lavner, 2008; Lee et al., 1997).

The general steps of SOLANU are shown in Figure 3-6, where the details of the epoch extraction and the derivation of the STM-MFCC and STM-LMPSC based speech boundaries are omitted, as they were provided in previous chapters. Epoch extraction was explained in Section 2.2.1, while the derivation of STM-MFCC and STM-LMPSC were provided in Section 3.3. The epochs of the speech signal were used by the PSOLA (Pitch Synchronous Overlap and Add) algorithm, covered in Section 2.2.2, to calculate the shift required to align the synthesized frames to minimize discontinuities in the signal. This procedure was identical to how existing speech rate modulation algorithms used epochs during alignment as was covered in Section 2.2. The difference between the existing designs and SOLANU was the method of calculating the stretch value at a given time.

The heart of SOLANU lies in the new method of deriving the stretch function, which utilized the identified speech boundaries. The stretch function is a modified triangular wave with a user defined range of $[1, \alpha_{max}]$, where α_{max} was the maximum stretch value. The frames identified as transition points based on the STM were assigned a stretch value of 1 to preserve the timing information. The frames on either side of this transient frame were given an incrementally increasing stretch value, corresponding to its distance from the closest transition point. In this thesis, the rate of increase in the stretch value was set so that it took 5 frames at a step size of 10 ms to start from a stretch factor of 1 and reach the maximum stretch factor of

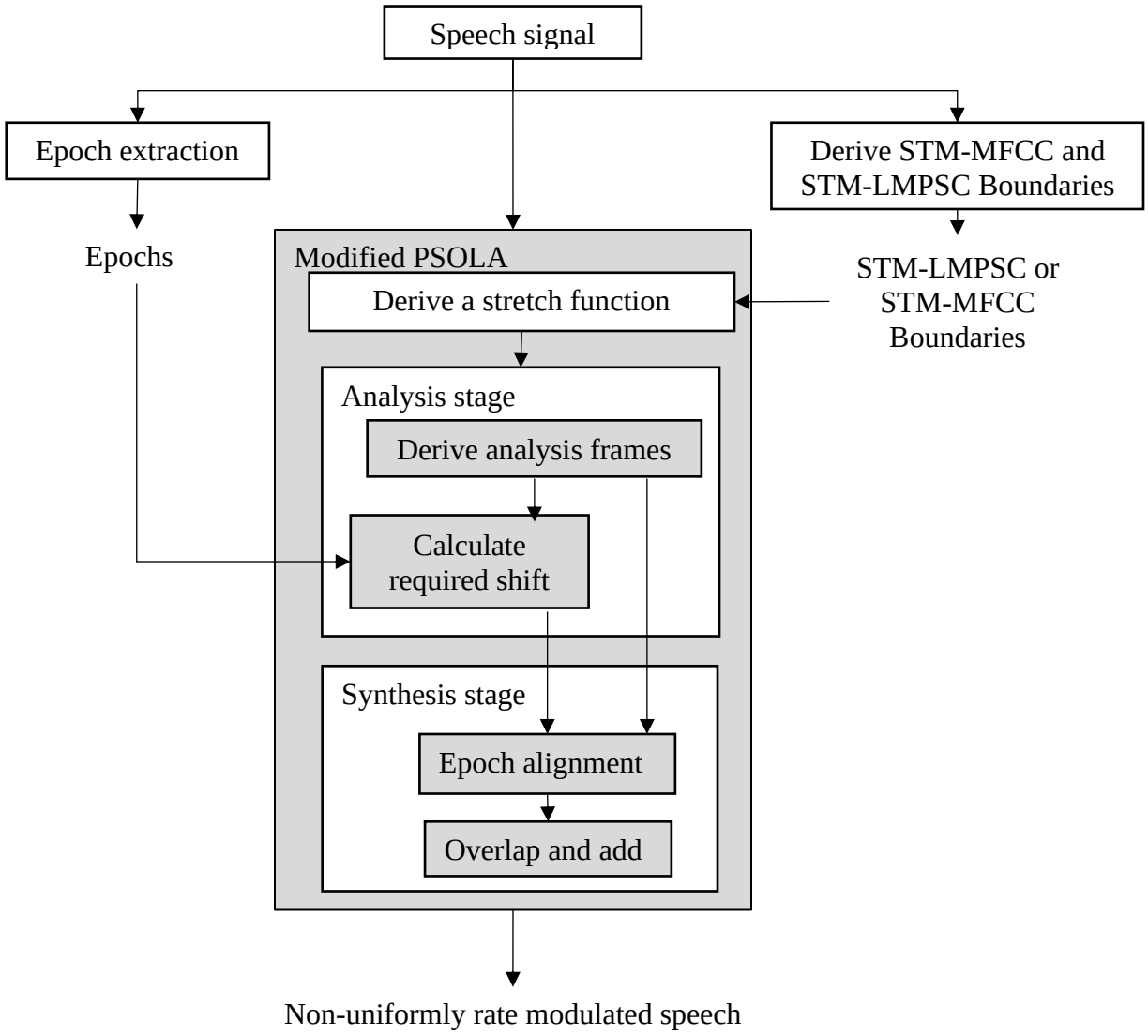


Figure 3-6: Schematic for deriving a non-uniformly rate modulated speech via SOLANU (SOLely Acoustic Non-Uniform) speech rate modulation. Details on epoch extraction and the derivation of the STM based speech boundaries is given in previous sections

α_{max}, or vice versa. In other words, there was a 50 ms span before and after an identified transient point where the speech decelerated, then accelerated. This was set so that the stretch function had adequate opportunities to reach its maximum value, as many of the speech boundariess were only separated by 100 ms or so. If there were not enough frames to allow for a full increase and decrease in the stretch factor before reaching the next speech boundary, the stretch function was set to start decreasing before reaching α_{max}. An example of the derived stretch function is shown in Figure 3-7, where the speech boundary points identified in yellow dots are preserved at a minimal stretch value, and the segments between these points are

incrementally slowed down with increasing stretch values. The use of MFCC to detect speech boundaries (Figure 3-7 (a)) yield similar, yet different stretch function as when LMPSC is used to detect the speech boundaries (Figure 3-7 (b)). The slope in the stretch function was chosen with consultation to the derived stretch function, ensuring maximum stretch value was reached in many occassions. However, future research is recommended to investigate the effects of the steepness of this increase and decrease in stretch value on the generated rate modulated speech.

The logic behind the design of SOLANU was that points where the greatest amount of spectral change is occurring in the speech, identified through the STM-MFCC or STM-LMPSC boundaries, should have its timing information preserved. This rationale is supported by several previous studies using a similar acoustic approach (Grofit & Lavner, 2008; Lee et al., 1997). In humans, the natural tendency is to slow down the vowels, but not so much the consonants within the words. This pattern is mimicked by assigning a stretch factor of 1 at these transients and incrementing the stretch factor with every subsequent frame located farther away from these identified boundaries. This yields a modified triangular wave as a stretch function with values

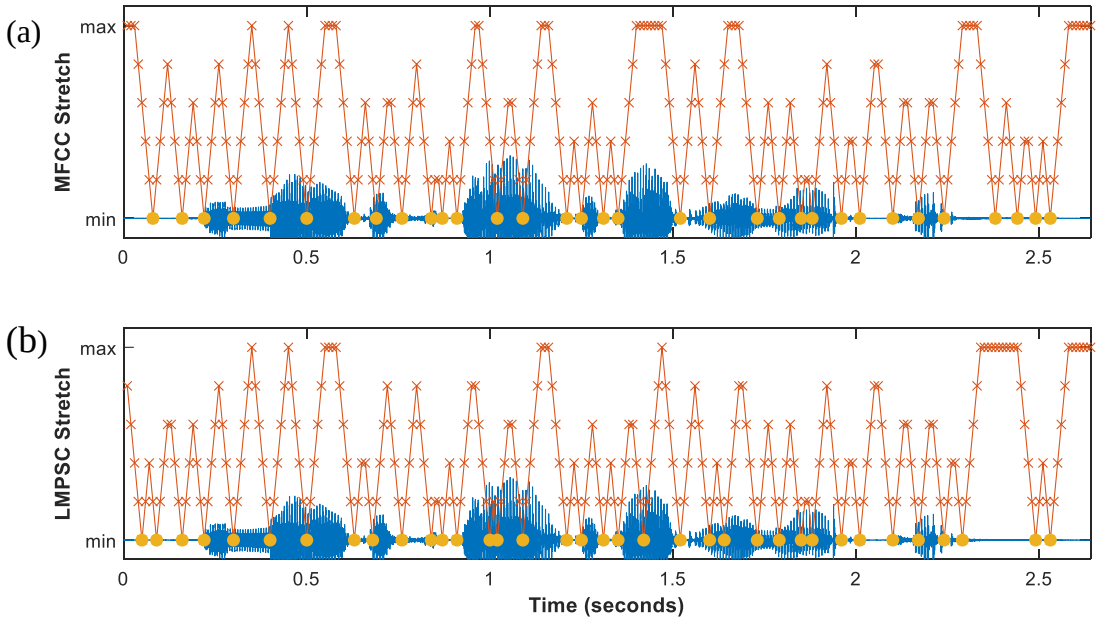


Figure 3-7: Stretch function (red line) derived based on speech boundaries (yellow dots) identified from STM-MFCC (above) and STM-LMPSC (below). Red x's indicate the actual stretch value applied to each frame, while the original speech signal is shown in blue. The timing information is preserved with a minimum stretch value at the speech boundaries, while the speech is slowed down with an increase in stretch value between any two given speech boundary points. MFCC and LMPSC variants of the stretch function were anticipated to be similar in many of the identified speech boundary locations, but slightly varying due to the differences in the cepstral and spectral coefficients.

within the range of $[1, \alpha_{max}]$. In this thesis, two variants of SOLANU are shown. One is associated with the STM-MFCC speech boundaries (Figure 3-7 (a)), while the other is associated with the STM-LMPSC speech boundaries (Figure 3-7 (b)). These variants were anticipated to produce similar but slightly different stretch functions due to the differences between MFCC and LMPSC.

SOLANU has some advantages and differences over existing non-uniform speech rate modulation algorithms. A major advantage of SOLANU is that it is capable of transforming speech without prior knowledge of consonant and vowel locations. It also does not require training of a vowel detector or classifier, eliminating the need for a large speech corpus with appropriate labels, or the need for specialized linguistics knowledge to generate such a database.

SOLANU is different from the two similar approaches in the method of identifying the spectral transition points. One existing approach uses cross-correlation in the MFCC (Grofit & Lavner, 2008), while the other existing approach uses another similar cepstral parameter (Lee et al., 1997). SOLANU uses STM of MFCC or LMPSC.

Another difference between SOLANU and the existing methods, lies in how the stretch function changes with time. In the simplest approach, a constant stretch value was used in the regions between the transients (Lee et al., 1997). In the other existing approach, a sawtooth type stretch function was used. Every time a spectral transition point was identified, the stretch function quickly increased, then gradually decreased until the next transient (Grofit & Lavner, 2008). In comparison, SOLANU generated a modified triangular wave as its stretch function. Whether one approach is better than the others, due to these differences in the method of identifying speech transients or shape of the stretch function adopted, is not known. This element was also not a research objective in this thesis, and will not be investigated further.

The literature review identified gaps in research, to which this thesis responded with new metrics and a novel non-uniform speech rate modulation algorithm. However, empirical studies are required to support the validity and reliability of these metrics and algorithms. Specifically, how SRBIM performance compares to existing measures of speech intelligibility in quantifying speech corrupted by noise. Similarly, how these metrics differ when measuring intelligibility of speech that has been enhanced by popular methods such as noise reduction, is an area of interest. Additionally, how the modified metrics as well as SRBIM perform when analyzing speech that

has been rate modulated is also unknown. Whether SOLANU performs differently than the existing uniform and non-uniform methods of speech rate modulation is also a study of interest.

Since there is a lack of studies on objectively measuring speech rate modulation type enhancement, it is desirable to test these research questions in a contained environment through incremental changes. In other words, there is a need for empirical research to investigate these questions in a controlled manner. The next section on research strategy will outline the general experimental flow taken to answer the questions above, and the rationale for choosing such a design.

3.5 Summary of Theory of Proposed Algorithms

This chapter has described modifications that can be implemented on existing metrics to make them capable of handling rate modulated speech, as well as the rationales and step-by-step derivation of the new SRBIM and SOLANU algorithms. The next chapter will provide the research methodology used to empirically validate these new algorithms.

4 Research Methods

The aim of this chapter is illustrate the experimental design and rationale in evaluating the performance of the newly proposed Speech Rate Based Intelligibility Metrics (SRBIM). The metrics' behaviours will be compared to existing speech intelligibility measures on a speech database containing noisy and clean speech in its original format, and the same speech stimuli that has been enhanced in one of two ways. One group of speech enhancement algorithms will improve the speech while preserving the rate of speech. In other words, this group of speech enhancement does not involve rate modulation. In this thesis, noise reduction algorithms were chosen for their popularity in literature as a general speech enhancement method. The second approach to speech enhancement modulates the rate of the speech. Speech intelligibility will be evaluated through a series of simulations involving the original noisy speech stimuli, the signals that have been processed through noise reduction, and signals that have been enhanced through both existing, and the proposed SOLely Acoustic Non-Uniform (SOLANU) speech rate modulation algorithms. This chapter will show the design of the experiments, and the rationale for the choices, as well as modifications and details required for replicating the simulations.

4.1 Research Strategy

This section will introduce the general design and rationale of the simulation experiments undertaken. This thesis implemented a series of simulation experiments that incrementally built upon the previous simulation. The speech stimuli used in the simulations contained speech with and without noise (Hu & Loizou, 2007). These speech stimuli were either tested as is, or enhanced through various algorithms before testing (Demol et al., 2004; ETSI, 2007; Islam, 2020; Lu & Loizou, 2008; Rudresh et al., 2018). The speech intelligibility metrics of each speech stimuli was calculated using the three existing measures of ESTOI (Extended Short-Time Objective Intelligibility), STOI (Short-Time Objective Intelligibility) and SNR (signal to noise ratio) loss, and the proposed SRBIM measures, MFCC-SRBIM (Mel Frequency Cepstral Coefficient type Speech Rate Based Intelligibility Metric) and LMPSC-SRBIM (Log Mel Power Spectral Coefficient type Speech Rate Based Intelligibility Metric). The performances of these metrics were compared with one another through correlation coefficients and scatter plots, as well as through metric agreement in a mock experiment judging noise reduction algorithm performance.

Experimentation through simulation was chosen in this thesis for several reasons. While several studies claimed that slowing speech down resulted in increased intelligibility as judged by human subjects, no previous study measured speech intelligibility of rate modulated speech through objective measures. There was a need to test the new SRBIM under controlled conditions where slight differences in the speech stimuli, such as type or level of noise, were possible for incremental comparisons.

The testing of the metrics using noisy, but otherwise unprocessed speech, was the focus of the first simulation. This speech stimuli consisted of thirty clean speech files, with thirty-two noise variants each, composed of eight different noise types, and four levels of noise. The details of this database are provided in Section 4.2.1. With such speech stimuli, the hypothesis was that noisy speech is less intelligible than clean speech, and this change should be indicated through the metrics. Additionally, incrementally louder noise was anticipated to further reduce speech intelligibility. This obvious setup provided a platform for comparing existing speech intelligibility metrics to SRBIM in a highly controlled environment, where the relative intelligibilities could be generally predicted. By first testing the speech intelligibility of simply noisy speech, this simulation avoided the speech intelligibility complications caused by additional processing. Furthermore, the metric behaviours observed under the first simulation can assist in interpreting the increases or decreases in the metric values after speech enhancement had been applied in later experiments. The metrics were compared in pairs, by analyzing through correlation coefficients and scatter behaviour. The scatter plots were included to clarify the type of relationship between the two metrics being compared.

The second simulation tested the three existing and two new metrics using speech stimuli that has been enhanced through noise reduction algorithms that involved neither speech rate modulation, nor other changes in the timing of speech features. Noise reduction was chosen for its popularity in speech enhancement. The same speech stimuli as above was first enhanced through noise reduction, then evaluated through the five speech intelligibility metrics. As discussed earlier, the metrics are compared using correlation coefficients and scatter plots to measure the strength and nature of the relationship. In this thesis, three different noise reduction algorithms were implemented, which are described in Chapter 4.2.2. The metrics were also used to evaluate the effectiveness of the different noise reduction algorithms, adding another dimension of comparison in addition to the correlation coefficients.

Speech rate modulation type enhancements were introduced in the next simulation, where the speech stimuli were processed by an existing uniform speech rate modulation algorithm, a slightly modified existing non-uniform speech rate modulation algorithm, and the two versions of the newly proposed SOLANU speech rate modulation algorithm, called SOLANU-MFCC and SOLANU-LMPSC. Similar to SRBIM, the SOLANU variants differed on the type of speech parameters used to detect the speech boundaries. The speech intelligibility metrics were similarly calculated and compared with each other via correlation and scatter plots. Some subjective observations were also included in the analysis.

The incremental comparison, which was provided as the rationale of choosing simulations in this thesis involving different levels of noise, also extends to the series of simulations. For example, SRBIM was proposed as a new speech intelligibility metric designed specifically for rate modulated speech, which had never been objectively measured before. Additionally, since SRBIM was a new metric, it had never been tested with speech enhanced through more popular methods, such as noise reduction. As such, before evaluating SRBIM performance with rate modulated speech, this thesis implemented a simulation testing SRBIM performance in speech enhanced through these traditional approaches. By comparing SRBIM behaviour to known speech intelligibility metrics under noisy and noise reduced speech, it is possible to postulate the observations made in the completely novel situation involving rate modulated speech.

Simulations also have the advantage of repeatability in the results. This is because computer simulations of objective measures output the same value given the same stimuli. Unlike human listeners, it is not affected by fatigue or language fluency. Additionally, objective measures do not remember past stimuli, allowing the same speech to be repeatedly presented under different noise levels or after applying different enhancement algorithms. This allows easy comparison between the various conditions, which is difficult with human subjects who remember previously presented speech stimuli. In short, humans can guess words based on both the current stimuli and previous knowledge, if the stimuli contained the same target speech, only differing in noise type, noise level or enhancement method.

This section gave the general experimental design of simulation experiments to incrementally test speech intelligibility across different controlled conditions. The next section will give the specifics of the methodology used in collecting the data.

4.2 Data Collection

Data collection in the simulations refers to the process of preparing the appropriate speech stimuli and calculating the three existing and two newly proposed speech intelligibility metrics. All simulations were carried out in MATLAB and the details required for replicating the experiment will be clarified in this section. Section 4.2.1 provides a description of the speech stimuli used, and the rationale for choosing this database. The three noise reduction algorithms used in the simulations are explained in Section 4.2.2. The stretch factors tested with the existing uniform and existing non-uniform method of speech rate modulation are covered in Section 4.2.3, while the parameters used in the new SOLANU algorithms are described in Section 4.2.4. Section 4.2.5 provides the details of the parameters required for SRBIM calculations, as well as the method of choosing the remaining two parameters required for peak picking. There was also a slight modification introduced into one of the existing metrics as it was attuned to quantifying only noise reduced speech. It was expanded to also handle noisy and rate modulated speech, which is covered in Section 4.2.6, along with information on where to find the relevant code. The final Section 4.2.7 gives the general flow of data collection, starting with the generation of the relevant speech stimuli, leading to the calculation of speech intelligibility with the five metrics.

4.2.1 Speech Stimuli

The speech corpus used in this experiment was the NOIZEUS database (Hu & Loizou, 2007). This database consists of thirty phonetically balanced IEEE sentences (IEEE Subcommittee on Subjective Measurements, 1969), recorded in a soundproof booth and professional recording equipment (Hu & Loizou, 2007). Three male and three female speakers produced five sentences each, covering all phonemes in the American English language with the thirty speech files (Hu & Loizou, 2007). The speech database was provided at a sampling frequency of 8 kHz (Hu & Loizou, 2007). The database is available to researchers free of charge from the original website (Loizou, n.d.-b), or alternatively, from a GitHub repository duplicated by another researcher (“NOIZEUS speech corpus,” 2018).

Apart from the original clean speech, NOIZEUS provides the noisy version of the speech containing eight types of real world noise at four different noise levels to produce thirty-two variants of each speech signal (Hu & Loizou, 2007). The noise sounds were from the AURORA database, and included airport, babble, car, exhibition, restaurant, station, street and train type noises (Hirsch & Pearce, 2000). Both babble type noise and restaurant type noise contained

multiple people talking in the background, but restaurant type noise could be described as being livelier. These noisy speech signals were provided at signal to noise ratios (SNR) of 15 dB, 10 dB, 5 dB and 0 dB (Hu & Loizou, 2007), which can be conceptualized as the difference between the speech and corresponding noise signals described in dB (IEEE Subcommittee on Subjective Measurements, 1969). Typically, the ratio of the powers of the signals are taken (ETSI, 2007), but other measures such as intensity or energy of the signals will also suffice (Yost, 2000). Mathematically, the SNR ratio is calculated as:

$$SNR_{ratio} = \frac{P_{speech}}{P_{noise}} \quad \textbf{(4-1)}$$

where P_{speech} and P_{noise} are the powers of the speech and noise, respectively. This SNR can then be converted into the dB scale with the following formula (Yost, 2000).

$$\begin{aligned} SNR_{dB} &= 10log_{10}(SNR_{ratio}) \\ &= 10log_{10}\left(\frac{P_{speech}}{P_{noise}}\right) \end{aligned} \quad \textbf{(4-2)}$$

As such, a SNR of 15 dB means that the power (or intensity, or energy) of the speech signal is 31.62 times that of the noise signal. A full list of the ratios of the powers corresponding to the four levels of noise used in the simulations are shown in Table 4-1.

Other than the fact that NOIZEUS provides phonetically balanced speech files with various noise types and levels, the database was chosen for this thesis for two main reasons. First, it serves well as a common basis for comparison across different research labs (Hu & Loizou, 2007) as it has been used in various studies on speech enhancement (Dhivya, Senthamizh, & Suresh, 2016; Y. Ma & Nishihara, 2014; Patil & Shirbahadurkar, 2018), as well as those evaluating new speech metrics (Hines, Skoglund, Kokaram, & Harte, 2015, 2013; Jassim & Zilany, 2016). Using a common NOIZEUS database in evaluating the validity of

Table 4-1: SNR in dB and its corresponding ratio between speech and noise signal powers

SNR_{dB}	$\frac{P_{speech}}{P_{noise}}$
15 dB	31.62 : 1
10 dB	10 : 1
5 dB	3.16 : 1
0 dB	1 : 1

speech rate measures as speech intelligibility metrics will make the research easily extendable by other research teams in the future.

Second, the database contained thirty-two variants of each speech stimuli, composed of eight types of noise incremented at four levels of noise in addition to its clean version. This allows for incremental comparison of the speech intelligibility metrics, highlighted as a major advantage in simulation type experiments. NOIZEUS allows for this comparison across varying levels and types of noise, as well as an aggregate analysis involving several stimuli groups.

As mentioned, the speech stimuli were provided in its original format for the first simulation. With subsequent simulations, the speech was enhanced before the metrics were calculated. The next few sections will cover noise reduction and speech rate modulation, the two types of speech enhancement techniques applied to the speech stimuli.

4.2.2 Noise Reduction

Three types of noise reduction algorithms were used to prepare three variants of the NOIZEUS database for speech enhanced through common methods that did not involve rate modulation. The three types of noise reduction algorithms were the geometric approach to spectral subtraction, the probabilistic geometric approach to spectral subtraction, and the two-stage mel-warped Wiener filter approach. Spectral subtraction and Wiener filter type noise reduction algorithms were chosen for this thesis as they are considered to be the two classical approaches in noise reduction (Heute, 2006). The probabilistic geometric approach was included in the simulations to allow investigation of the behaviour of the proposed metrics when measuring speech enhanced through two similar noise reduction algorithms.

Spectral subtraction assumes that additive noise can be cancelled by simply subtracting it from the overall signal in the frequency domain (Heute, 2006). The geometric approach to spectral subtraction (Lu & Loizou, 2008) and its variant, the probabilistic geometric approach to spectral subtraction (Islam, Shahnaz, Zhu, & Ahmad, 2018), were used in this thesis. The MATLAB code for the geometric approach to spectral subtraction is available from (Loizou, n.d.-a). The MATLAB code for the probabilistic geometric approach was provided by (Islam, 2020).

On the other hand, Wiener filter method attempts to approximate the clean speech signal through minimum mean squared error (Heute, 2006). The two-stage mel-warped Wiener filter based design was taken from the technical report issued by the European Telecommunications

Standards Institute (ETSI, 2007), which I translated into MATLAB code. The algorithm converts the noisy signal into the frequency domain, estimates the noise spectrum with the help of a voice activity detector, calculates filter coefficients using this noise spectrum estimate, scales the filter coefficients to the mel scale, and implements a mel-warped inverse discrete cosine transform to apply the filter to the noisy speech signal in the time domain (ETSI, 2007). The denoised speech signal can then be processed through the filter once more, where noise reduction is more dynamically controlled the second time around (ETSI, 2007).

To produce the noise reduced versions of the speech signals in MATLAB, first, the original NOIZEUS speech signals were converted from .wav files into arrays representing the digital signal. The noisy speech signal arrays were then processed through a noise reduction algorithm and the output was power normalized while its maximum negative and positive amplitudes were simultaneously adjusted to be within the range of -1 and 1. This prevents clipping of the signal when writing out an audio file, and balances the power of the output signal based on the power in the original speech signal. These adjustments do not affect the objective intelligibility metrics as the existing metrics normalize the parameters before comparison across time. On the other hand, clipping would affect the existing metric outcomes as the signal characteristics in time would be lost at these clipped portions. Normalization would not be able to recover the lost information with clipping, but can accommodate for differences in scaling brought about by the adjustments. The noise reduced and normalized signal was saved to an appropriately named folder corresponding to the three different noise reduction algorithms. All thirty-two noisy speech variants, as well as the original clean version of the thirty speech signals in NOIZEUS were processed through the three noise reduction algorithms. Metric calculations of noise reduced speech would start by taking in this noise reduced speech signal saved to the computer, rather than starting with the original speech signals and applying noise reduction every time.

4.2.3 Existing Speech Rate Modulation Algorithms

For speech rate modulation type speech enhancement, one existing uniform speech rate modulation, one existing non-uniform speech rate modulation, and two variants of the newly proposed SOLANU speech rate modulation algorithms were used. I coded all algorithms in MATLAB for generating speech rate modulated versions of the NOIZEUS signals.

For all three speech rate modulation algorithms, the PSOLA (Pitch Synchronous Overlap and Add) component, which was explained in Section 2.2.2, was responsible for time scaling the speech signals according to a given stretch factor or function. Uniform speech rate modulation used a constant stretch factor, while the non-uniform approaches had a separate component that derived a time-varying stretch function, which was used by PSOLA to appropriately rate modulate the speech. All speech rate modulation algorithms used 50 % overlap between neighbouring synthesis frames, with window lengths of 30 ms, and used a linear decay function when weighting the importance of the previously synthesized signal and the current frame to be added during the final overlap and addition phase. The 50 % frame overlap was taken from the original paper (Rudresh et al., 2018), while the 30 ms window length was chosen as it produced slightly better sounding output compared to the 20 ms window length used in the original paper. The linear function was chosen as it was one of the suggested choices in the original paper as well (Rudresh et al., 2018).

PSOLA also relied upon epoch extraction for proper alignment of the synthesized frames. Generally, default values were used of epoch extraction (Murty & Yegnanarayana, 2008). However, to better extract the epoch locations, two modifications were made. One modification was to incorporate an estimate of the pitch period within the code rather than use a predefined value, which was a modification that was suggested in a later paper published by the same research group that proposed the epoch extraction method (Yegnanarayana & Gangashetty, 2011). In the original paper, the slope or trend in the signal was removed by working on a signal segment, 10 ms in length (Murty & Yegnanarayana, 2008). This 10 ms segment length was a crude estimate of the pitch period of the voice. However, a later publication used autocorrelation to better estimate the average pitch period (Yegnanarayana & Gangashetty, 2011). This search was limited between 10 ms and 30 ms and sought for the maximum correlation between the current signal with its time shifted version. The location of maximum correlation implied that the signal was synced in phase, allowing the user to estimate its pitch period.

However, even with this improved pitch period estimation, the slope in the signal was not completely removed. This removal of the trend is crucial in epoch identification, as the final step in epoch extraction is dependent on the location of the positive zero crossings, or where the negative signal turns into a positive value (Murty & Yegnanarayana, 2008). If the trend is not completely removed, the signal only crosses the zero boundary over a very small portion of the

whole speech. As such, the second modification that I implemented forced trend removal by calculating the overall slope remaining in the signal and removing it through simple subtraction. While PSOLA and its associated epoch extraction was integrated in all speech rate modulation algorithms used in this thesis, there were parameters and operations specific to certain modulation algorithms.

For uniform speech rate modulation, a constant stretch factor was required. The stretch factors of 1.0, 1.2, 1.4, 1.6, 1.8 and 2.0 times the original rate were used. The 1.0 factor was chosen to investigate whether distortions were introduced by the algorithm even without actual rate modulation. The 2.0 factor was the maximum stretch allowed using a synthesis frame overlap of 50 %. The stretch factors were incremented by 0.2 to give six stretch factors in uniform speech rate modulation based on PSOLA.

On the other hand, the non-uniform speech rate modulation algorithms required a time-varying stretch function. The stretch value would be confined to the user specified limits of a minimum and maximum stretch factor. In this thesis, only the maximum stretch factor was manipulated. The minimum stretch factor was kept at 1.0 to conserve the timing information of the transients in SOLANU, which was the theoretical basis from which the algorithm was proposed. The same range of stretch factors were used for the existing non-uniform speech rate modulation algorithm for consistency across the two non-uniform approaches. The maximum stretch was manipulated between 1.0, 1.2, 1.4, 1.6, 1.8 and 2.0 times the original speech rate to allow comparison of the non-uniform approaches to the uniform approach.

The existing non-uniform speech rate modulation algorithm used in this thesis was similar to SOLANU as it was separated into a component that derived the stretch function, and another component that actually stretched the speech signal (Demol et al., 2004). However, this algorithm used WSOLA (Waveform Similarity Overlap and Add), another variant of the Overlap and Add (OLA) family, for rate modulating the signal. In this thesis, this was modified into PSOLA for two reasons. First, there were no previous studies that had objectively measured speech intelligibility of rate modulated speech. Thus, it was desirable to implement similar rate modulation techniques where comparison across the different algorithms was possible. If both the stretch function derivation and the signal stretching components were varied between the existing model and SOLANU, it may be difficult to attribute differences in the metric performance to either the stretch calculation or the actual stretching. By standardizing the

stretching operation to the PSOLA algorithm, the intelligibility metrics derived from the different modulation algorithms could be compared.

Another reason for using PSOLA instead of WSOLA lies in the similarity of the two algorithms. Both algorithms align the signal during synthesis and work in the time domain. This similarity allows substitution of WSOLA with PSOLA without many anticipated problems. In addition to this major change in the stretching operation, minor modifications were also introduced as described in the next section.

As mentioned in Chapter 2, many of the existing non-uniform speech rate modulation algorithms are poorly documented, making it difficult to replicate. There was enough detail to replicate the general flow of the stretch derivation of the algorithm mentioned above, including details on the parameters such as window length of 20 ms and step size of 5 ms, but the thresholds used, and the stretch factors used were not well explained (Demol et al., 2004). For example, the algorithm requires three threshold values while deriving the stretch function. This includes the energy threshold to classify pauses, the threshold for the transition rate of the root mean square value of the speech signal for determining plosives, and the threshold associated with the Average Magnitude Difference Function (AMDF), which was required to detect periodicity of the signal to determine vowels, consonants and phone transitions. Of these three values, only the energy threshold was specified at 0.008 for a speech signal normalized to an amplitude of [-1, 1] (Demol et al., 2004). From the figures in the paper, the transition rate threshold and AMDF threshold were inferred to be 0.6 and 0.5, respectively, but the rationale for these values were unclear (Demol et al., 2004). In this thesis, the default parameters were used under the assumption that the theory of the original paper, although not provided, was sound, and that the same threshold values can be used with the NOIZEUS database used in the simulations.

For the existing non-uniform approach introduced in Section 2.3, the δ_x values and α value which dictate the stretch for a given speech segment was also not documented in detail. It was given that the final stretch factor was calculated as $\alpha_{final} = \alpha \div \delta_x$, and the relative relationship of the parameters was specified to be $\delta_1 < \delta_2 < \delta_3 < \delta_4 < 1$ for the pauses (δ_1), vowels (δ_2), consonants (δ_3), phone transitions (δ_4) and plosives (1) (Demol et al., 2004). However, how these values were to be determined was not stated in existing papers, nor was it specified whether these values were to be uniformly or non-uniformly incremented (Demol et al., 2004). In this thesis, the values were manipulated according to the user defined minimum and

maximum stretch values mentioned above, where the final stretch factor was $\alpha_{final} = \alpha \div \delta_x$, where $\alpha = 1$ was used. Furthermore, the parameters were adjusted so that the final stretch factors were evenly incremented between the five different speech classifications. For example, if the maximum stretch was specified to be 2.0, the plosive stretch was 1.0, the phase transition stretch was 1.25, the consonant stretch was 1.5, the vowel stretch was 1.75, and the pause stretch was 2.0. Figure 4-1 shows a sample speech signal and its corresponding δ_x and α_{final} values with a stretch range from 1.0 to 2.0.

4.2.4 SOLANU

The other non-uniform speech rate modulation algorithm used in this thesis was the newly proposed SOLANU (SOLely Acoustic Non-Uniform) algorithm, available in its SOLANU-MFCC (Mel Frequency Cepstral Coefficient) variant as well as its SOLANU-LMPSC (Log Mel Power Spectral Coefficient) variant. The general flow of SOLANU was explained in Section 3.4, and PSOLA and its associated epoch extraction method was explained in Section 2.2, so this section will go into specific parameters used in deriving the triangular wave like stretch function.

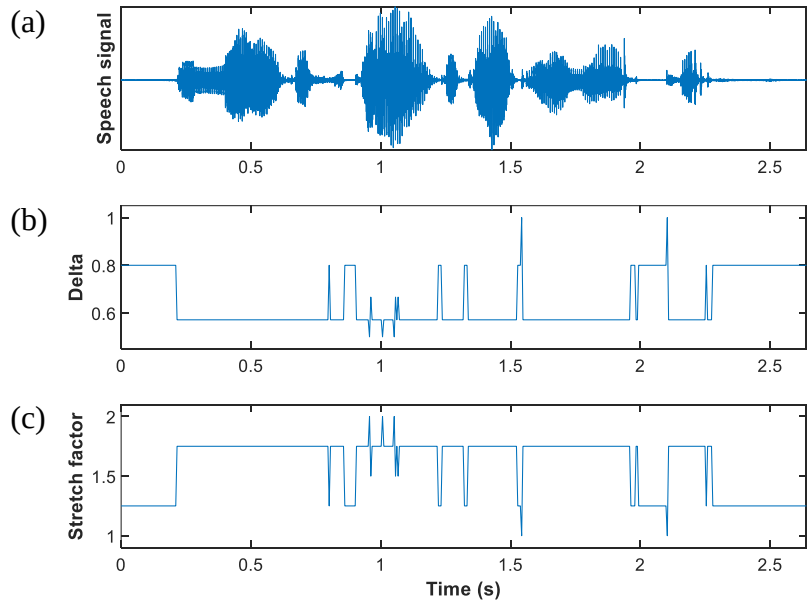


Figure 4-1: A plot of a (a) sample speech signal, which has been assigned a (b) delta value according to speech segment classification. This is converted into a (c) stretch value to be applied during the existing non-uniform speech rate modulation algorithm

As mentioned before, this thesis used two variants of SOLANU, based on the spikes in the STM (Spectral Transition Measure) signal derived from MFCC and LMPSC. The general flow of deriving these values were shown in Section 3.3, but parameters such as window length and step sizes used in the simulations were not explicitly stated.

There were five parameters required for calculating MFCC and LMPSC and one parameter associated with obtaining the STM. The parameters for coefficient calculation were window length of analysis, the step size that these windows were shifted by, the type of windowing applied to these frames, the number of triangular mel filter banks used and the number of cepstral coefficients to be obtained. In the simulations, the Hamming type window of length 32 ms and step size 10 ms were used in the analysis. This was the most common combination of window type, window length and step size used in previous studies, including the study that used STM-MFCC for speech rate estimation (Aharonson et al., 2017), the original STM paper (Furui, 1986) and various studies comparing different methods of MFCC calculations (Ganchev, 2011). For the triangular filter banks, twenty filters were used based on the narrowband model (Ganchev, 2011). The number of filters designated the number of spectral coefficients and the common range seemed to be twenty to twenty-four (Ganchev, 2011). On the other hand, the number of cepstral coefficients derived varied a bit more from study to study detailed in the literature. The general recommendation was a number that was less than the number of filters used, and somewhere within the range of thirteen to nineteen (Ganchev, 2011), but there were also cases where fewer coefficient numbers also sufficed for measuring speech rate (Aharonson et al., 2017). As such, for the simulations, a value on the lower end of the spectrum was chosen at thirteen cepstral coefficients. In analyzing the STM signal, the default parameters of two frames on either side of the current frame of analysis was used, as similar window lengths and step sizes of 30 ms and 10 ms were used in the STM study (Furui, 1986). This STM parameter was kept at two frames as the original STM study found that the combination of these parameters could successfully measure and separate transients in the speech that were important for speech perception (Furui, 1986).

The two parameters which were not specified in literature were the window length and the multiplicative threshold factor used in the median based adaptive threshold peak picking algorithm. These parameters were empirically optimized by looking for the maximum magnitude

of the correlation between existing metrics and SRBIM (Speech Rate Based Intelligibility Metrics), which also used the parameters specified above in detecting speech boundaries.

4.2.5 SRBIM

SOLANU and SRBIM use many of the same calculations, starting with obtaining the speech parameters, converting it into its STM value, and identifying the peaks in the STM. The STM peaks were classified as speech boundaries, and the location of these transients were used in SOLANU to derive a stretch function. On the other hand, the total number of speech boundaries within the speech segment was used in calculating SRBIM. As such, SRBIM also used the same five parameters for coefficient calculations, one parameter for STM estimation, and two parameters for detecting the peaks.

Since SRBIM performance was to be measured by comparison with existing speech intelligibility metrics, the two unknown peak picking parameters of window length and multiplicative threshold factor were chosen by maximizing the magnitude of the correlation between MFCC-SRBIM and LMPSC-SRBIM with existing metrics, using the original, unprocessed NOIZEUS database. During this optimization step, the window lengths of 30 ms, 50 ms, 70 ms, 90 ms, 110 ms, 130 ms, 150 ms, 170 ms, 190 ms and 210 ms were tested. Since the step size used was 10 ms, this meant that the median value was searched in a sample size ranging from 3 to 21 points for a selection of ten choices. The threshold value, on the other hand, was tested for values ranging from 0.2 to 3.0 in increments of 0.2 for a total of fifteen choices. A total of 150 combinations of peak picking parameters were used to calculate the two SRBIM. The correlation between each SRBIM with the three existing speech intelligibility metrics was calculated, and the threshold and window length combination that produced the largest magnitude in correlation that was consistent with expectations and consistent in behaviour across the three measures was chosen. Once optimized, the parameters were kept constant for the remaining simulations.

4.2.6 Additional Adjustments to Existing Speech Intelligibility Metrics

For measures of ESTOI (Extended Short-Time Objective Intelligibility) and STOI (Short-Time Objective Intelligibility), the default parameters were used. With SNR (signal to noise ratio) loss, the SNR range which was set to a default of [-3, 3] dB was readjusted to [-15, 15] dB, the range used in SII (Speech Intelligibility Index) calculations (J. Ma & Loizou, 2011). This modification was made as SNR loss was previously used to measure speech intelligibility of

noise reduced speech, where the SNR range was limited. For this thesis, SNR loss was used in measuring intelligibility of noisy speech, time scaled speech as well as noise reduced speech, which was anticipated to involve a larger SNR range. The MATLAB code for ESTOI and STOI were available from Sound Zone Tools, a GitHub repository of various MATLAB functions used in auditory processing ("Sound Zone Tools: Signal Processing Tools for MATLAB," 2017). This was because the web links for the free MATLAB implementation in the original papers were inaccessible at this time (Jensen & Taal, 2016; Taal et al., 2010). The MATLAB code for SNR loss was obtained from the researcher's website (Loizou, n.d.-a).

4.2.7 General Flow of Data Collection

Three sets of simulations were executed in this thesis, based on differences in the speech stimuli whose intelligibility was measured through five different metrics. The first simulation used the original NOIZEUS database with thirty speech signals, each given in its clean format, as well as thirty-two noise variants based on eight noise types at four noise levels. The second set of simulations used NOIZEUS speech signals that had been noise reduced. The three types of noise reduction algorithms generated three variants of the NOIZEUS database, all of which was measured with the three existing and two new speech intelligibility measures. The simulations were designed so that files would be saved at the end of each process. I programmed the MATLAB code such that one MATLAB function produced the relevantly enhanced speech database as a collection of .wav files. Three other MATLAB functions calculated each of the three existing speech intelligibility metrics for the whole database, while MFCC-SRBIM and LMPSC-SRBIM were calculated simultaneously, but through two successive functions. Each MATLAB function saved a .mat file which could be loaded by subsequent MATLAB functions for data analysis, which is covered in the next section.

4.3 Framework for Data Analysis

After the speech intelligibility metrics were calculated for all the prepared speech files, the five metrics were compared against each other through correlation coefficients and scatter plots. This was done for noisy speech, noise reduced speech as well as rate modulated speech. The enhanced speech was combined with the original noisy speech during correlation analysis. With the noise reduced speech, an additional analysis was presented, based on the mock comparison of the three noise reduction algorithms.

4.3.1 Correlation Between Speech Intelligibility Metrics

The Pearson correlation coefficients were calculated between the three existing and two newly proposed speech intelligibility measures derived during data collection. The aim was to evaluate the performance of the new metrics by comparing to existing metrics. A common approach taken when a new intelligibility metric was proposed was to find its correlation with the mean subjective ratings collected from human observers. For example, past studies have correlated new objective intelligibility metrics to the subjective scores of intelligibility or quality as rated by human listeners (J. Ma & Loizou, 2011; Taal et al., 2011). The correlation assumes linear dependence between the two metrics under comparison, and indicates the strength of the association between the two. This was then generalized as indicating the strength of the metric in objectively measuring speech intelligibility.

In this thesis, the correlation coefficients were calculated between all ten combinations of the five objective measures, rather than using subjective scores as the ultimate reference. This method was chosen because obtaining subjective rating for the NOIZEUS database would be difficult under the experimental design used in this thesis. As mentioned before, this database was excellent for incremental comparisons of speech stimuli that were masked by different types or levels of noise. However, since the same target speech stimuli are used repeatedly under different noise conditions, it is difficult to ask human subjects to identify the spoken words based solely on the presented material, and not on a prior presentation of a variant of the same stimuli. As such, SRBIM performance was evaluated indirectly through correlation with existing measures and compared against the degree of variation observed within ESTOI, STOI and SNR loss.

For example, calculating the correlation of the performance between MFCC-SRBIM and ESTOI, STOI or SNR loss gives a measure of the output similarity between the new measure and the three existing metrics. By comparing this correlation value to the correlation observed within ESTOI, STOI and SNR loss, it is possible to evaluate the SRBIM performance as a strength in its association with existing measures, as compared to the strength in the association between two existing and accepted measures. This can be visualized as an evaluation of SRBIM performance through indirect means. Since ESTOI, STOI and SNR loss are measures that are assumed to be correlated relatively well with subjective scores, by finding the correlation coefficient between MFCC-SRBIM or LMPSC-SRBIM to these existing measures, the validity of using SRBIM to

evaluate speech intelligibility can be indirectly investigated. Additionally, by comparing ESTOI, STOI and SNR loss performance among themselves, it is possible to see the strength of agreement that exists between various existing measures of intelligibility.

When calculating correlation coefficients, it is desirable to include samples with varying levels of speech intelligibility. This is because a more reliable association between the two metrics being compared can be obtained by assessing through a larger spectrum, rather than a small selection of samples. This can be visualized as the difference of obtaining the line of best fit through a collection of points that are scattered along the full range of x and y values which roughly represent a linear relationship, compared to finding the line of best fit through a cluster of points aggregating at one point on the same linear relationship. As such, all four noise levels of each noise type were treated collectively as one sample set. For the first simulation testing the original NOIZEUS dataset, the metrics were compared for each of the eight noise types, as well as on the aggregate or whole dataset. Analyzing these correlations would reveal whether metric correlations were significantly different for one type of noise and give an indication whether the lower correlation in the aggregate data was caused by this one noise type, or due to the independence of the linear relationships between the different noise types. If the latter were the case, each noise type wise correlations would be relatively high, while the aggregate data would be low in comparison. For simulations involving different speech enhancement techniques, these original noisy speech signals were included in the correlation calculations to contribute samples to the lower end of the intelligibility spectrum.

Error in the correlation was calculated by the leave-one-out approach through the thirty speech files. Namely, once the correlation coefficient was calculated using the full dataset appropriate for the selected noise type, the error was derived from the standard deviation of the correlations that selectively omitted a speech file from the set. For example, there were thirty speech files presented in its clean version and in its noisy versions. As such, thirty different correlation coefficients were calculated, each omitting one of the thirty speech signals. During this removal of the speech signals, its clean version, and all its noise variants were removed as a group. The standard deviation in this sample of thirty correlation coefficients were used as the error in the calculated correlation.

In addition to correlation coefficients, scatter plots are analyzed to observe the nature of the relationship between the two metrics being compared. Correlation coefficients assume

linearity between the two metrics and output the strength in association between the two metrics. As such, correlations by themselves can be misleading as to the strength of the relationship, when it is not linear. The analysis of the scatter plots will be able to confirm whether metric relationships are linear or nonlinear.

Apart from correlation coefficients and scatter plot analysis, for the noise reduction simulations, the metrics were also evaluated through a mock experiment. With this analysis, the noise reduction algorithms were evaluated on their effectiveness in improving speech intelligibility, as measured by improvements quantified in the five intelligibility metrics.

4.3.2 Mock Evaluation of Noise Reduction Algorithm Performance

For the noise reduction simulations, a mock experiment of evaluating the three noise reduction algorithms was carried out to add another dimension of analysis. This experiment assumed that all five metrics were valid in measuring speech intelligibility and compared the algorithm performance based on the improvement in the metrics from the original noisy speech. The metric differences were also normalized to the metric value calculated with the speech before enhancement for better comparison between different speech samples. By deriving this difference for all three noise reduction algorithms, the level of improvement caused by each of the algorithms could be visualized independently through the five metrics. A similar comparison for the clean speech signals before and after noisy reduction revealed the degree of distortion introduced through the noise reduction algorithm, assuming the clean signal had no noise at all.

These differences in the metrics were then compared across the three noise reduction algorithms through a series of paired t-tests. This was much like how an actual experiment, evaluating the effectiveness of three noise reduction algorithms, would progress. By momentarily assuming the validity of the metrics, this mock experiment allows the comparison in the behaviour of the metrics in relation to how they rank different noise reduction algorithms. Similar to the correlation coefficients, this agreement pattern between the new SRBIM and existing metrics can be compared to the agreement pattern within ESTOI, STOI and SNR loss. While correlations were better calculated with a dataset involving a large range of speech intelligibilities, this mock experiment could be isolated to each noise level as well as noise type. As such, the metric agreements were collected for the each of the thirty-two noise conditions, a slightly larger dataset that combined all noise levels within each of the eight noise types, or a

separate dataset where noise types were combined for analysis of each of the four noise levels, as well as the aggregate data consisting of a mix of noise types and noise levels.

SRBIM performance has been evaluated through a comparative study with existing metrics through correlation as well as a mock experiment in the noise reduction simulation. For evaluating the speech enhancement techniques involving rate modulation, the same comparisons were carried out, along with observations based on my subjective comments on the intelligibility of the rate modulated speech. This was included as it was anticipated that many of the modified existing speech intelligibility metric performances would deteriorate with non-uniformly rate modulated speech, due to the loss of frame alignment.

4.4 Summary of Research Methods

This chapter has thus described the details of the research methodology used in the simulations including the rationale for the chosen research strategy, the method of data collection starting with the appropriate speech stimuli, the required speech enhancements of noise reduction or rate modulation, and the five speech intelligibility metrics calculated. The method of analyzing the data based on correlation coefficients, scatter plots and a mock experiment involving the rating of three noise reduction algorithms was described. The next chapters will present the research findings from these simulations.

5 Simulation Results of Noisy Speech

The results obtained using the methodology described in Chapter 4 are presented over the next three chapters. This chapter presents the simulation findings obtained for the original speech with different levels of noise that have not processed through any enhancement algorithm. This includes the results obtained during the optimization of the peak picking parameters of SRBIM (Speech Rate Based Intelligibility Metrics). Chapter 6 describes the results obtained from speech that had been enhanced through the three types of noise reduction algorithms. Finally, Chapter 7 details the simulation findings pertaining to speech that had been enhanced through rate modulation, including existing uniform and non-uniform speech rate modulation methods and the two newly proposed SOLANU (SOLely Acoustic Non-Uniform) algorithms.

This chapter will start with the results of the parameter optimization for the peak picking algorithms required for calculating the MFCC-SRBIM (Mel Frequency Cepstral Coefficient type Speech Rate Based Intelligibility Metric) and the LMPSC-SRBIM (Log Mel Power Spectral Coefficient type Speech Rate Based Intelligibility Metric). The behaviour of these two proposed SRBIM will then be compared to the three existing speech intelligibility metrics of ESTOI (Extended Short-Time Objective Intelligibility), STOI (Short-Time Objective Intelligibility) and SNR (signal to noise ratio) loss. This comparison will be analyzed between pairs of metrics through correlation coefficients and scatter plots for noisy, but otherwise unprocessed, speech.

5.1 Optimization of Peak Picking Parameters

Calculations for MFCC-SRBIM and LMPSC-SRBIM require many user defined parameters. While many of these parameters were based on those used in previous papers calculating STM-MFCC (Spectral Transition Measure of Mel Frequency Cepstral Coefficients) (Aharonson et al., 2017; Furui, 1986; Ganchev, 2011; Young et al., 2009), the two parameters used in the peak picking algorithm were not provided in previous papers, and had to be derived empirically. The aim was to determine the combination of window length and threshold factor that produced the best correlation between MFCC-SRBIM and LMPSC-SRBIM with the three existing metrics of ESTOI, STOI and SNR loss. The speech database used was the original, unprocessed NOIZEUS database (Hu & Loizou, 2007) detailed in Chapter 4. For parameter optimization, all eight noise types were aggregated to construct a model that would work

relatively well in all noise types. The optimized parameters were used for the remainder of the simulations.

The peak picking process was optimized by varying the window length and the threshold factor. These parameters dictated the adaptive threshold used in determining the peaks in the STM signal. Specifically, the window length was the number of STM points to consider when calculating the median at any given point. The adaptive threshold was obtained by multiplying the determined median by the threshold factor. The STM signal must exceed the adaptive threshold to be classified as a peak. MFCC-SRBIM and LMPSC-SRBIM were optimized by comparing ten window lengths between 30 ms to 210 ms, and fifteen thresholds between 0.2 and 3.0, for a total of 150 variants. Each condition was then correlated with ESTOI, STOI and SNR loss metrics and evaluated in the strength of the correlation and the statistical significance of the relationship. SRBIM was first compared to ESTOI.

5.1.1 Correlation with ESTOI for Parameter Optimization

This section will present the findings on the strength and direction of the correlations between ESTOI and MFCC-SRBIM and LMPSC-SRBIM obtained through the 150 combinations of threshold and window lengths used in the peak picking algorithm.

The correlation between ESTOI and MFCC-SRBIM is visualized in Figure 5-1. The threshold value had an obvious effect of changing the directionality and strength of the correlation between MFCC-SRBIM and ESTOI, while the window length did not have a strong influence. Many of the weak correlations near zero were identified to have no statistical significance, noted by the absence of dots on the intersections of the mesh graph in Figure 5-1, and italicized values in Table 5-1.

Otherwise, correlations with moderately strong positive or negative strengths were considered to have statistical significance ($p < 0.001$). Of the correlations that were statistically significant, the strongest negative correlation occurred at threshold of 1.0, as seen by the large dip in Figure 5-1. The exact values are shown in Table 5-1, where a threshold factor of 1.0 produced correlations in the range of -0.620 ±0.006 to -0.694 ±0.005. Strongest positive correlations were observed at a threshold 0.2 with values ranging from 0.738 ±0.008 to 0.788 ±0.003, and at threshold 3.0, where positive correlations were in the range of 0.669 ±0.008 to 0.727 ±0.005. Logically, parameter optimization was a search for the combination of window lengths and thresholds that produced the largest magnitude in the correlations between ESTOI

and MFCC-SRBIM, combined with the expected directionality. Thus, threshold of 1.0 was chosen for its large negative correlation, and window length 90 ms was noted to have the largest magnitude in the correlation, as highlighted in Table 5-1. The reason for choosing negative correlation will be elucidated in the paragraphs below.

The choice of the optimized threshold value was partially explained through a scatter plot of the MFCC-SRBIM against ESTOI. Figure 5-2 (a) shows a plot of MFCC-SRBIM against ESTOI at a threshold of 0.2 for all speech signals contained in NOIZEUS. It was evident that many of the 0 dB SNR signals (purple diamonds) had reached the lower limit of MFCC-SRBIM, as they had clustered to a value near 0 for several speech signals with a measured ESTOI between 0.3 and 0.4. This meant that threshold 0.2 would be unable to distinguish between varying degrees of intelligibility at the lower extreme. In comparison, threshold 1.0 in Figure 5-2 (b) and threshold 3.0 in Figure 5-2 (c) show the plots to be relatively evenly spaced, with margins for quantifying speech intelligibility outside the range where the majority of the points are currently placed. The major difference between these two remaining threshold candidates was the directionality of the correlations.

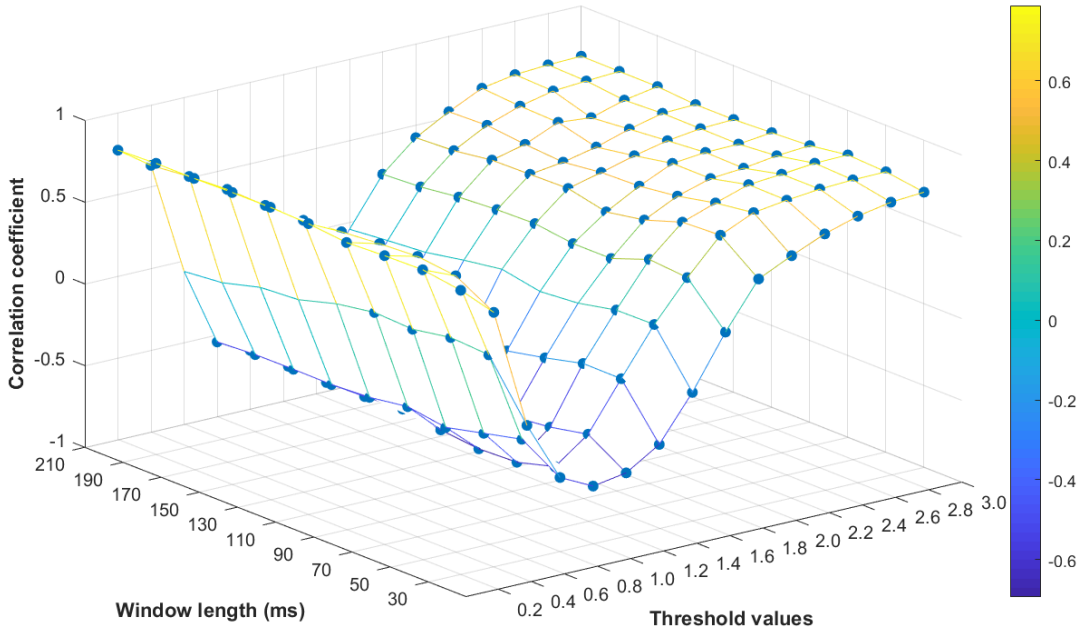


Figure 5-1: Correlation coefficient between ESTOI and MFCC-SRBIM calculated using different combinations of window length and threshold values in picking peaks. Dots signify points of statistical significance ($p < 0.001$). Threshold had a larger effect on correlation than window length

Table 5-1: Correlation between ESTOI and MFCC-SRBIM calculated using different combinations of threshold and window length ($p < 0.001$). Threshold 1.0 (highlighted row) showed the most negative correlations

Thres-hold	Window length (ms)									
	30	50	70	90	110	130	150	170	190	210
0.2	0.738 ±0.008	0.772 ±0.005	0.767 ±0.004	0.756 ±0.003	0.780 ±0.003	0.788 ±0.003	0.787 ±0.003	0.781 ±0.003	0.782 ±0.003	0.771 ±0.003
0.4	0.556 ±0.006	0.688 ±0.004	0.716 ±0.003	0.702 ±0.004	0.683 ±0.005	0.661 ±0.006	0.661 ±0.005	0.668 ±0.005	0.655 ±0.005	0.631 ±0.005
0.6	-0.182 ±0.009	0.158 ±0.008	0.172 ±0.008	0.131 ±0.009	0.143 ±0.009	*0.104 ±0.008*	*0.037 ±0.008*	*0.021 ±0.008*	*-0.042 ±0.009*	*-0.06 ±0.01*
0.8	-0.548 ±0.009	-0.405 ±0.008	-0.460 ±0.008	-0.516 ±0.007	-0.481 ±0.008	-0.515 ±0.006	-0.529 ±0.007	-0.527 ±0.007	-0.529 ±0.006	-0.543 ±0.006
1.0	-0.647 ±0.006	-0.620 ±0.006	-0.685 ±0.005	-0.694 ±0.005	-0.670 ±0.006	-0.634 ±0.006	-0.649 ±0.005	-0.654 ±0.005	-0.647 ±0.005	-0.641 ±0.007
1.2	-0.613 ±0.005	-0.468 ±0.006	-0.517 ±0.006	-0.542 ±0.005	-0.506 ±0.007	-0.530 ±0.007	-0.534 ±0.006	-0.52 ±0.01	-0.539 ±0.006	-0.557 ±0.007
1.4	-0.486 ±0.006	-0.176 ±0.007	-0.174 ±0.007	-0.231 ±0.007	-0.280 ±0.006	-0.282 ±0.007	-0.291 ±0.005	-0.321 ±0.005	-0.311 ±0.007	-0.332 ±0.006
1.6	-0.214 ±0.007	0.109 ±0.008	0.106 ±0.008	*0.057 ±0.008*	*0.04 ±0.01*	*0.081 ±0.009*	*0.045 ±0.009*	*0.001 ±0.008*	*-0.019 ±0.008*	*-0.041 ±0.009*
1.8	0.108 ±0.007	0.349 ±0.009	0.369 ±0.009	0.283 ±0.009	0.28 ±0.01	0.313 ±0.008	0.31 ±0.01	0.292 ±0.008	0.263 ±0.008	0.249 ±0.008
2.0	0.386 ±0.005	0.562 ±0.006	0.550 ±0.008	0.472 ±0.008	0.430 ±0.007	0.482 ±0.007	0.48 ±0.01	0.471 ±0.007	0.438 ±0.006	0.426 ±0.008
2.2	0.481 ±0.006	0.653 ±0.004	0.621 ±0.004	0.571 ±0.007	0.541 ±0.009	0.533 ±0.01	0.50 0.02	0.52 ±0.01	0.546 ±0.007	0.538 ±0.009
2.4	0.568 ±0.006	0.682 ±0.007	0.686 ±0.006	0.632 ±0.008	0.633 ±0.006	0.588 ±0.008	0.55 ±0.02	0.609 ±0.008	0.617 ±0.007	0.635 ±0.005
2.6	0.630 ±0.006	0.703 ±0.006	0.703 ±0.006	0.648 ±0.006	0.653 ±0.005	0.633 ±0.006	0.61 ±0.01	0.59 ±0.01	0.662 ±0.006	0.671 ±0.006
2.8	0.668 ±0.005	0.713 ±0.005	0.723 ±0.005	0.677 ±0.005	0.657 ±0.005	0.660 ±0.006	0.662 ±0.009	0.65 ±0.01	0.677 ±0.007	0.684 ±0.006
3.0	0.683 ±0.006	0.715 ±0.005	0.727 ±0.005	0.693 ±0.006	0.680 ±0.006	0.669 ±0.006	0.669 ±0.008	0.688 ±0.006	0.688 ±0.006	0.690 ±0.007

The negative direction in the correlations was the desired behaviour in MFCC-SRBIM as compared to ESTOI. Since STM-MFCC, which is the basis of MFCC-SRBIM, detects the transition points of speech, faster changing speech would have more boundaries within a given time. In addition to the boundaries from the original speech, the presence of fluctuating noise

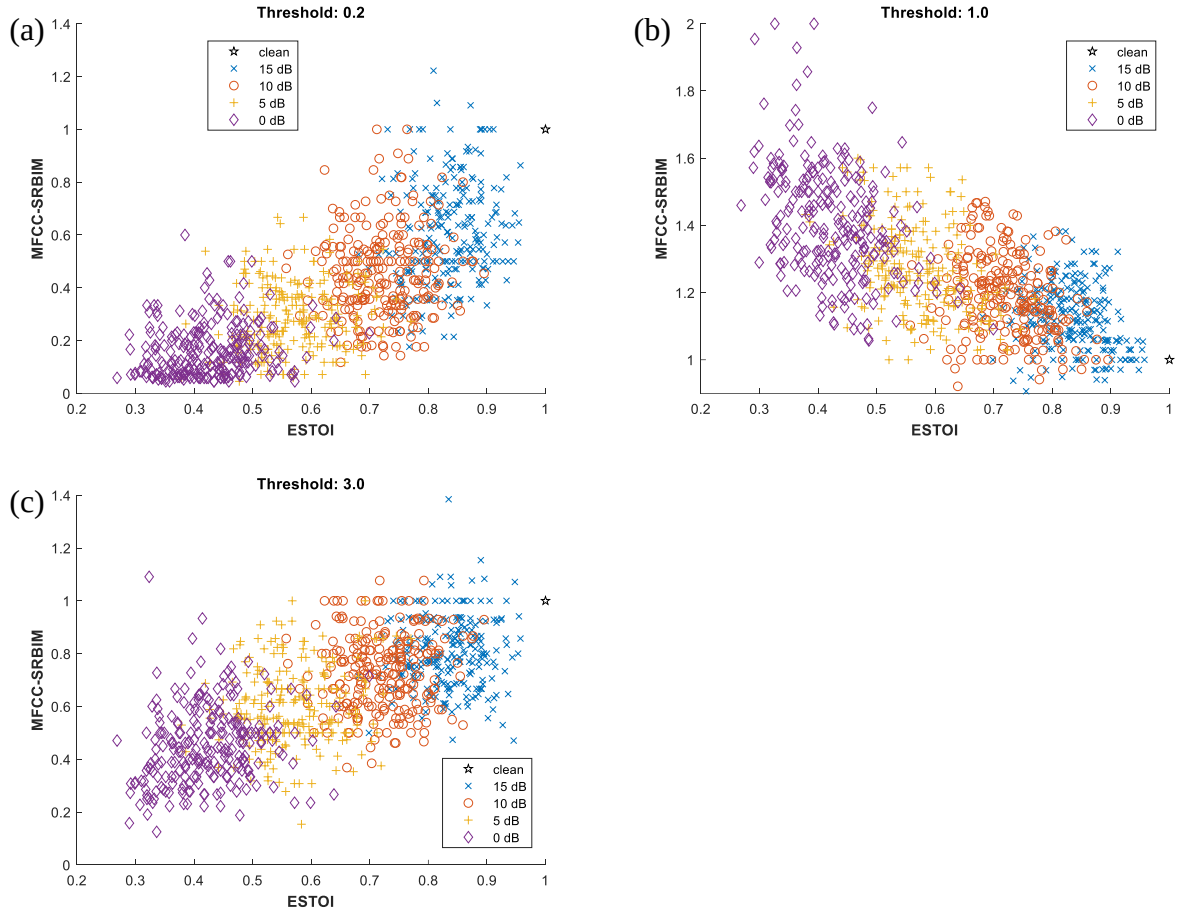


Figure 5-2: MFCC-SRBIM plotted against ESTOI using window length 90 ms and thresholds at (a) 0.2, (b) 1.0 and (c) 3.0. The different colours and shapes of the plots indicate the noise levels in SNR and the clean signal with no noise. Threshold 1.0 shows negative correlation between MFCC-SRBIM and ESTOI

was anticipated to contribute more boundaries to the STM signal. As such, noisy speech was expected to have higher STM-MFCC speech rates compared to the clean speech, resulting in a normalized rate or MFCC-SRBIM greater than 1. Conversely, ESTOI metrics were expected to decrease with additional noise. Therefore, ESTOI and MFCC-SRBIM were expected to have negative correlation, suggesting threshold 1.0 to be the most appropriate choice. For window length, 90 ms was chosen as this gave the most negative correlation of -0.694 ±0.005 (dark shaded cell in Table 5-1).

The choice of negative correlation can be further supported by Figure 5-3. This figure shows the STM-MFCC signal (blue line) of one sample speech and its corresponding peaks (red lines and circles) as detected for each scenario differing in threshold used and the amount of noise present. Figure 5-3 (a) and (b) used threshold 0.2 (top row), Figure 5-3 (c) and (d) used

threshold 1.0 (middle row), and Figure 5-3 (e) and (f) used threshold 1.0 (bottom row). The effects of the thresholds are most visible in the resulting adaptive threshold, indicated by the black line. The left column was based on a speech signal with no noise, while the right column contained noise at SNR of 5 dB. From these graphs, the effects of threshold and noise level on STM-MFCC rates could be observed.

As a general observation, Figure 5-3 shows that more peaks were detected under the optimal threshold 1.0 (c and d), in comparison to thresholds 0.2 (a and b) or 3.0 (e and f). This was due to the fact that increasing the multiplicative threshold factor increased the actual adaptive threshold that the STM signal must surpass in order to be considered a peak. Threshold 3.0 (bottom row) decreased the frequency at which STM (blue line) exceeded this adaptive threshold (black line), which ultimately resulted in fewer peaks and a lower calculated rate.

Conversely, for threshold 0.2, a lower multiplicative threshold factor, the occurrence of the STM signal descending below this adaptive threshold decreased. Since the peak picking algorithm was programmed to choose only one peak from each segment, defined by regions where the STM was below the adaptive threshold, threshold 0.2 (top row) also resulted in lower STM-MFCC rates in comparison to threshold 1.0. This mechanism explains how STM-MFCC rate changed as a function of threshold. However, the direction of the correlation with ESTOI needed to consider the effects of noise and the resulting changes in MFCC-SRBIM.

The direction of the correlation changed with threshold because noise influenced the STM-MFCC signal, and its interactions with the peak picking algorithm. MFCC-SRBIM is the ratio of the speech rates between the noisy speech and the clean speech. A negative correlation between MFCC-SRBIM and ESTOI at threshold 1.0 dictated that increased levels of noise resulted in higher normalized rates, or greater MFCC-SRBIM. In other words, higher STM-MFCC rates were observed for noisy speech in comparison to clean. The blue lines in Figure 5-3 show the STM-MFCC signal for clean speech (left column) and speech masked by babble noise at 5 dB SNR (right column). The change in the overall scale is not important as the peak picking algorithm generates an adaptive threshold based on the median of the given signal. However, it was observed that the noisy STM-MFCC changed its amplitude more frequently as compared to the clean speech. With threshold 1.0, Figure 5-3 (d), these frequent amplitude changes were selected as peaks, resulting in an increase in MFCC-SRBIM with increasing noise, and a negative correlation with ESTOI.

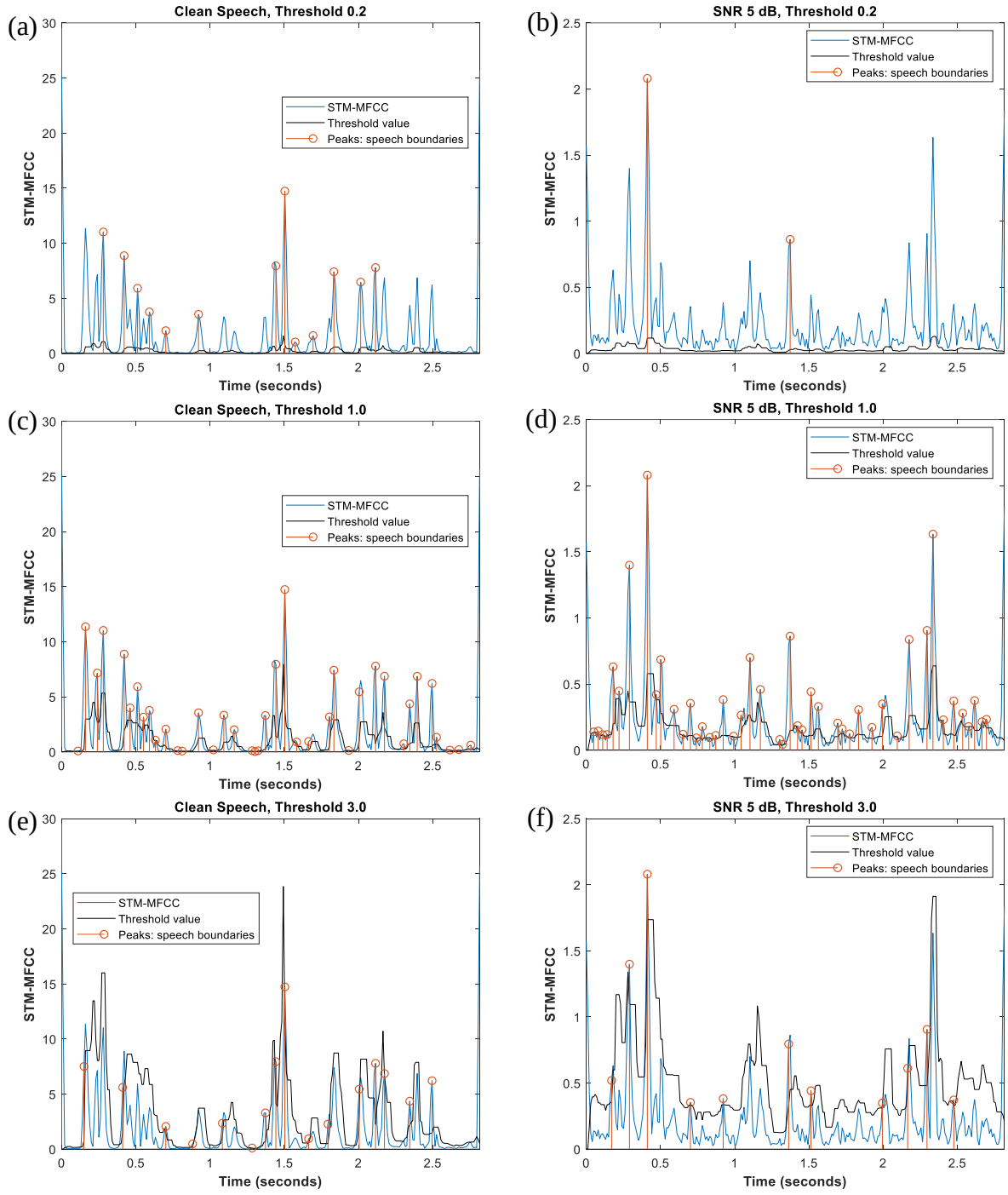


Figure 5-3: STM-MFCC peaks picked for one speech signal at three different thresholds for its clean speech and noisy speech using window length of 90 ms. (a) Clean speech at threshold 0.2, (b) SNR 5 dB noisy speech at threshold 0.2, (c) clean speech at threshold 1.0, (d) SNR 5 dB noisy speech at threshold 1.0, (e) clean speech at threshold 3.0 and (f) SNR 5 dB noisy speech at threshold 3.0. Threshold 1.0 detects more peaks with increasing noise.

On the other hand, a positive correlation with ESTOI observed at thresholds 0.2 and 3.0 dictated that the additive noise decreased the number of speech boundaries detected, as demonstrated in Figure 5-3 (b) in comparison to (a), and Figure 5-3 (f) in comparison to (e). The noise caused more frequent peaks in STM and reduced the segments where the STM signal was near zero. These near zero segments were where the STM-MFCC signal at threshold 0.2 most often dipped below the adaptive threshold, defining speech segments where one peak was chosen from each segment. Without these near zero segments, the peak picking algorithm was forced to select a peak from a small number of long segments. This resulted in a decrease in normalized rate with noise, leading to a positive correlation at threshold 0.2.

In addition to this overall increase in the STM floor value, the noise reduced the prominence of the peaks. This affected the peaks chosen at threshold 3.0, where it become even more difficult for the STM-MFCC signal to exceed the adaptive threshold via these smaller, albeit more frequent, peaks. This also led to a decrease in normalized rate with noise, resulting in a positive correlation at threshold 3.0.

As such, additive noise reduced MFCC-SRBIM detected with threshold 0.2 and 3.0, resulting in a positive correlation with ESTOI. Thus, threshold 1.0 produced the desired negative correlations between ESTOI and MFCC-SRBIM. The window length had a smaller effect, but the largest magnitude in the correlations were observed at 90 ms for MFCC-SRBIM.

From a theoretical standpoint, the direction of the correlation between MFCC-SRBIM and ESTOI could be either positive or negative. ESTOI was a speech intelligibility measure that could theoretically range from [-1, 1], but its output was expected to settle mostly in the range of [0, 1], with lower values reflecting lower intelligibility. SRBIM, on the other hand, had a range of $(0, \infty)$, where the clean signals were normalized to a value of 1. This meant that in theory, the direction of the correlation between SRBIM and ESTOI could be positive if less intelligible speech had SRBIM values less than 1, which was the case with thresholds of 0.2 and 3.0, or negative if less intelligible speech had SRBIM values greater than 1, as was the case with threshold of 1.0.

However, from a practical standpoint, a negative correlation was desired, as greater levels of noise was expected to produce more data and more peaks, rather than dwindling down the number of peaks to ultimately result in 1 peak for the whole signal. This was further supported by the behaviour of the scatter plots in Figure 5-2 where threshold 1.0 showed margins for

measuring speech intelligibility outside the range given by the noisy database. Additionally, Figure 5-3 showed that the presence of noise, reduced the number of peaks detected at thresholds 0.2 and 3.0, resulting in a oversimplification of the signals and a loss of information.

While MFCC-SRBIM was not constrained by an upper bound for this thesis, it is common practice in other existing metrics to cap the upper limit to a predetermined value. For investigative purposes, as SRBIM was a newly proposed metric, the values were left as is in this thesis, but an upper bound may be used for future studies.

Similar results to MFCC-SRBIM were observed in LMPSC-SRBIM under different peak picking window lengths and threshold values. As shown in Figure 5-4, correlations between ESTOI and LMPSC-SRBIM were negative for threshold 1.0 and its neighboring values, but positive for all other thresholds, and window length had very little effect on the calculated correlations. There were a couple points with no statistical significance as indicated by the lack of dots in Figure 5-4 and italicized values in Table 5-2, where the calculated correlations were near zero. All other parameter combinations showed statistical significance ($p < 0.001$). However, unlike MFCC-SRBIM, correlations between ESTOI and LMPSC-SRBIM did observe a slight decline in magnitude from threshold 0.4 to 0.2.

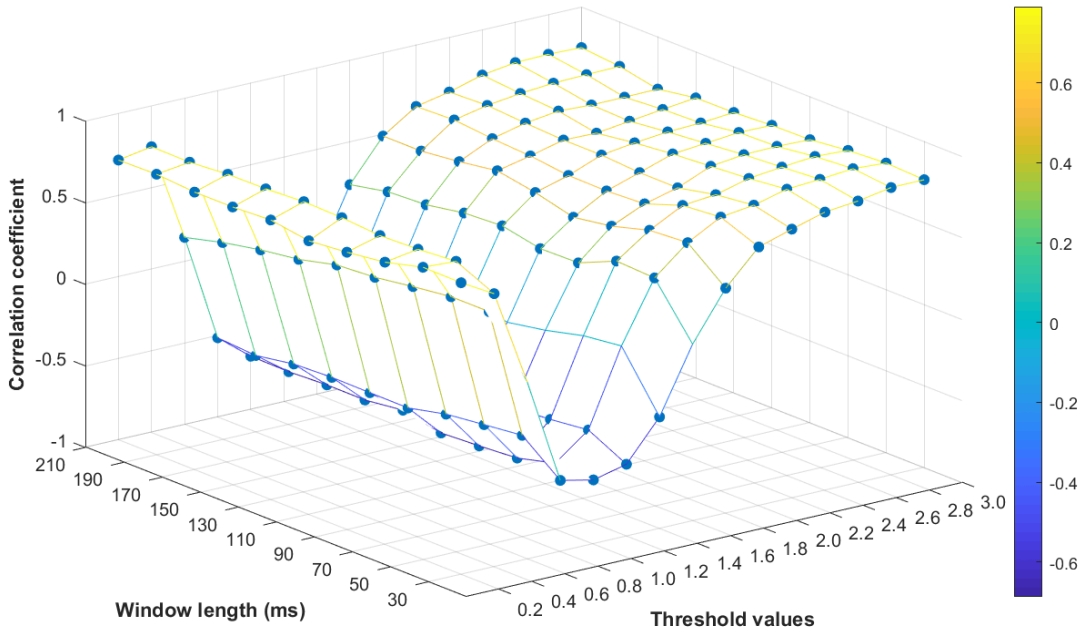


Figure 5-4: Correlation coefficient between ESTOI and LMPSC-SRBIM calculated using different combinations of window length and threshold values in picking peaks. Dots signify points of statistical significance ($p < 0.001$). Threshold had a larger effect on correlation than window length

Table 5-2: Correlation between ESTOI and LMPSC-SRBIM calculated using different combinations of threshold and window length ($p < 0.001$). Threshold 1.0 (highlighted row) showed the most negative correlations

	Window length (ms)									
Thres-hold	30	50	70	90	110	130	150	170	190	210
0.2	0.788 ±0.003	0.791 ±0.003	0.730 ±0.004	0.700 ±0.004	0.678 ±0.003	0.714 ±0.003	0.702 ±0.004	0.702 ±0.005	0.720 ±0.004	0.715 ±0.005
0.4	0.674 ±0.006	0.781 ±0.003	0.763 ±0.007	0.755 ±0.004	0.773 ±0.004	0.780 ±0.003	0.767 ±0.003	0.765 ±0.003	0.747 ±0.003	0.750 ±0.004
0.6	*0.085* *±0.007*	0.425 ±0.006	0.425 ±0.007	0.395 ±0.008	0.361 ±0.008	0.345 ±0.009	0.288 ±0.009	0.254 ±0.009	0.208 ±0.008	0.147 ±0.008
0.8	-0.565 ±0.008	-0.380 ±0.009	-0.406 ±0.008	-0.434 ±0.009	-0.487 ±0.007	-0.485 ±0.007	-0.483 ±0.007	-0.493 ±0.007	-0.532 ±0.007	-0.515 ±0.008
1.0	-0.609 ±0.006	-0.598 ±0.006	-0.658 ±0.005	-0.675 ±0.005	-0.687 ±0.005	-0.640 ±0.006	-0.669 ±0.005	-0.667 ±0.005	-0.678 ±0.005	-0.675 ±0.006
1.2	-0.560 ±0.005	-0.441 ±0.007	-0.468 ±0.007	-0.485 ±0.007	-0.457 ±0.008	-0.459 ±0.009	-0.486 ±0.008	-0.488 ±0.008	-0.492 ±0.008	-0.491 ±0.007
1.4	-0.316 ±0.007	*0.027* *±0.007*	*-0.002* *±0.008*	*-0.060* *±0.008*	*-0.102* *±0.008*	*-0.053* *±0.007*	-0.111 ±0.007	-0.143 ±0.005	-0.165 ±0.006	-0.187 ±0.006
1.6	*0.083* *±0.007*	0.398 ±0.009	0.409 ±0.007	0.307 ±0.008	0.302 ±0.007	0.350 ±0.007	0.338 ±0.007	0.296 ±0.006	0.285 ±0.008	0.237 ±0.008
1.8	0.380 ±0.008	0.567 ±0.008	0.551 ±0.008	0.485 ±0.007	0.45 ±0.01	0.51 ±0.01	0.550 ±0.008	0.514 ±0.006	0.487 ±0.007	0.488 ±0.007
2.0	0.585 ±0.006	0.677 ±0.006	0.657 ±0.006	0.575 ±0.008	0.56 ±0.01	0.566 ±0.009	0.59 ±0.01	0.608 ±0.007	0.631 ±0.005	0.622 ±0.006
2.2	0.648 ±0.007	0.703 ±0.006	0.676 ±0.006	0.626 ±0.009	0.628 ±0.007	0.644 ±0.006	0.645 ±0.007	0.657 ±0.006	0.674 ±0.006	0.664 ±0.006
2.4	0.705 ±0.006	0.754 ±0.004	0.728 ±0.005	0.711 ±0.007	0.696 ±0.007	0.672 ±0.009	0.64 ±0.01	0.684 ±0.007	0.700 ±0.007	0.721 ±0.006
2.6	0.726 ±0.004	0.768 ±0.004	0.751 ±0.005	0.736 ±0.005	0.714 ±0.006	0.692 ±0.009	0.67 ±0.01	0.674 ±0.008	0.716 ±0.007	0.751 ±0.004
2.8	0.762 ±0.004	0.786 ±0.004	0.770 ±0.005	0.740 ±0.005	0.734 ±0.005	0.700 ±0.009	0.69 ±0.01	0.675 ±0.009	0.722 ±0.006	0.755 ±0.005
3.0	0.763 ±0.004	0.774 ±0.004	0.760 ±0.005	0.734 ±0.007	0.731 ±0.006	0.718 ±0.007	0.715 ±0.009	0.68 ±0.01	0.726 ±0.007	0.749 ±0.006

Aggregate noisy speech data plotted in terms of LMPSC-SRBIM against their respective ESTOI metrics also showed similar results as MFCC-SRBIM. Figure 5-5 (a) at threshold 0.2 shows that the normalized metric of LMPSC-SRBIM reached a limit at the lower intelligibilities of SNR 0 dB (purple diamonds), making differences at this lower bound difficult to discriminate.

On the other hand, threshold 1.0 in Figure 5-5 (b), and threshold 3.0 in Figure 5-5 (c), demonstrated potential in quantifying speech with lower or higher intelligibilities than the ones currently tested.

Since LMPSC-SRBIM was also expected to have a negative correlation with ESTOI, threshold 1.0 was chosen to be the default value. As seen in the highlighted row in Table 5-2, this threshold offered the strongest negative correlation out of all the tested thresholds. Window size of 110 ms had the greatest negative correlation at -0.687 ±0.005 and was chosen as the optimized value (dark highlight in Table 5-2). The mechanism for the negative correlation at threshold 1.0 and positive correlations at thresholds 0.2 and 3.0 was the same as that explained for STM-MFCC. Increasing levels of noise affected the STM-LMPSC signal such that fewer peaks were detected at thresholds 0.2 and 3.0 and more peaks were detected at threshold 1.0 with increasing levels of noise.

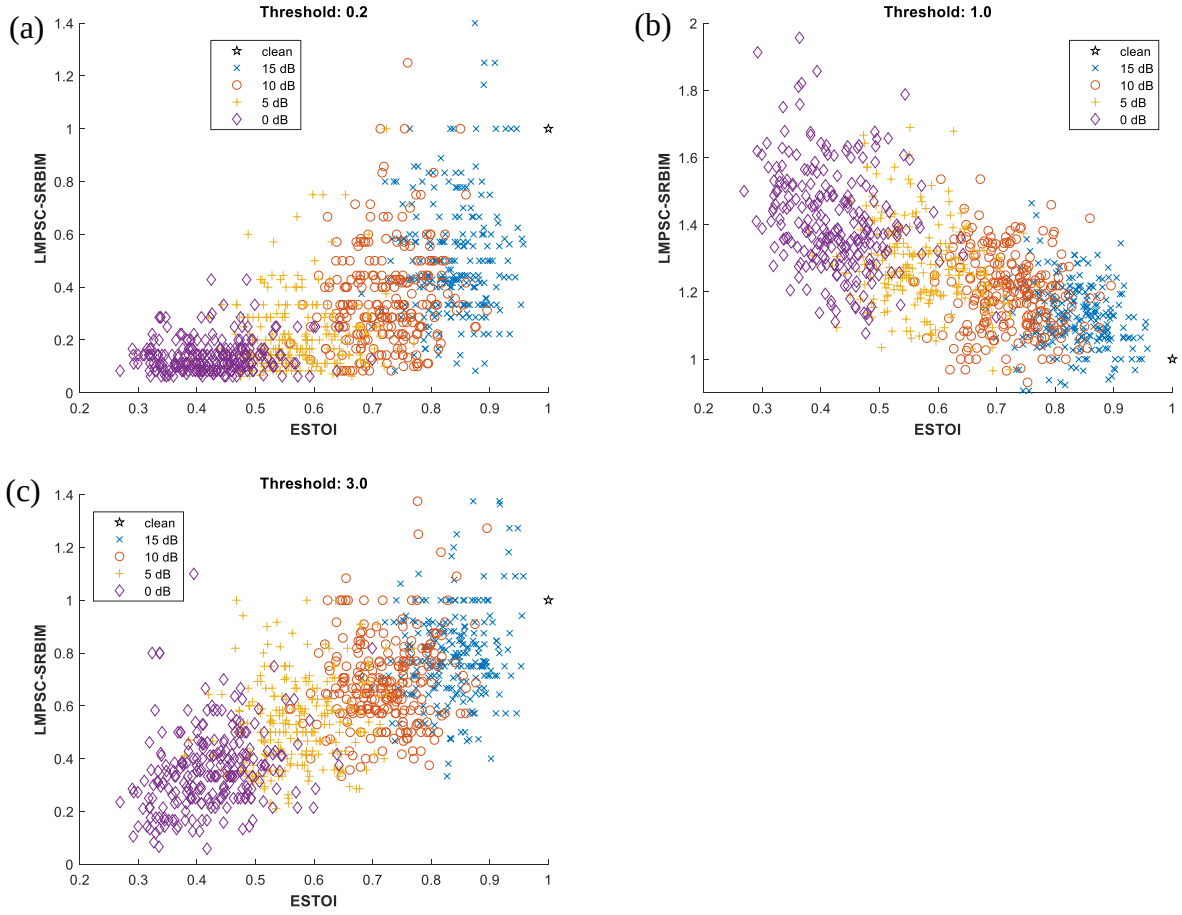


Figure 5-5: LMPSC-SRBIM plotted against ESTOI using window length 110 ms and thresholds at (a) 0.2, (b) 1.0 and (c) 3.0. The different colours and shapes of the plots indicate the noise levels in SNR and the clean signal with no noise. Threshold 1.0 shows negative correlation between LMPSC-SRBIM and ESTOI

By finding the largest negative correlations between SRBIM and ESTOI, the optimized value of the threshold parameter was suggested to be 1.0, while the optimal window length were slightly different at 90 ms for MFCC-SRBIM and 110 ms for LMPSC-SRBIM. The correlations between SRBIM and STOI were investigated next to determine if the same optimized parameter values were obtained regardless of the existing metric SRBIM was compared against.

5.1.2 Correlation with STOI for Parameter Optimization

The magnitude and directionality of the correlation between STOI and SRBIM were similar to that observed between ESTOI and SRBIM. Namely, threshold 1.0 had the greatest negative correlation, and window length had a significantly smaller effect.

Figure 5-6 and Figure 5-7 show the plots of correlation between STOI and MFCC-SRBIM and LMPSC-SRBIM, respectively, for the different combinations of window lengths and threshold values tested in the peak picking algorithm. While the optimal threshold, as determined through correlations with STOI, was the same as that from ESTOI, correlation magnitudes observed for different window lengths showed a slight difference for LMPSC-SRBIM. The complete table containing the correlation values is contained in Appendix A for MFCC-SRBIM, and Appendix B for LMPSC-SRBIM.

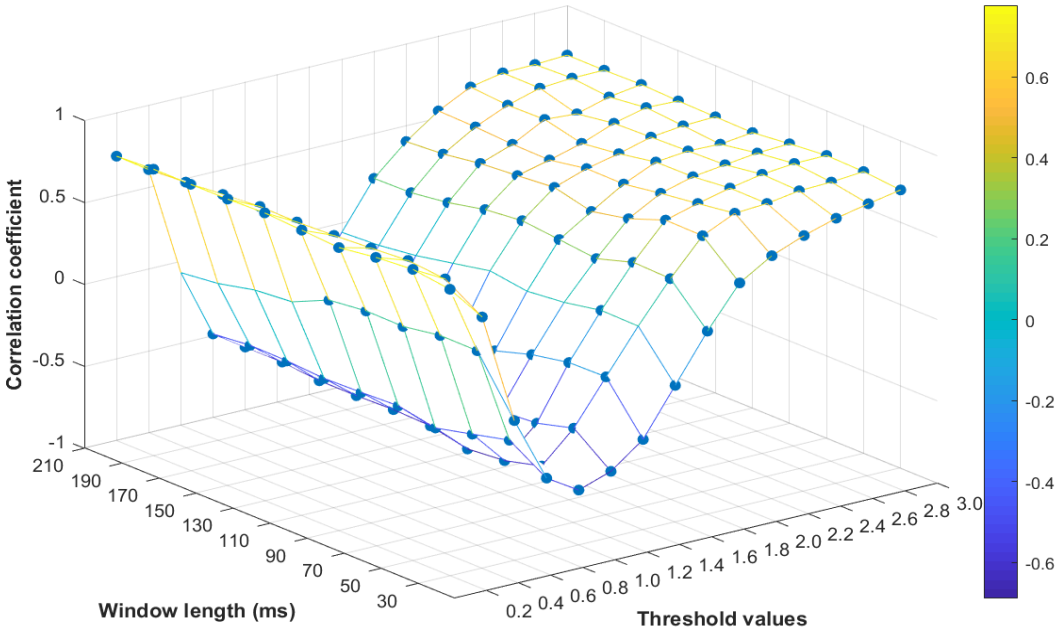


Figure 5-6: Correlation coefficient between STOI and MFCC-SRBIM calculated using different combinations of window length and threshold values in picking peaks. Dots signify points of statistical significance ($p < 0.001$). Threshold had a larger effect on correlation than window length

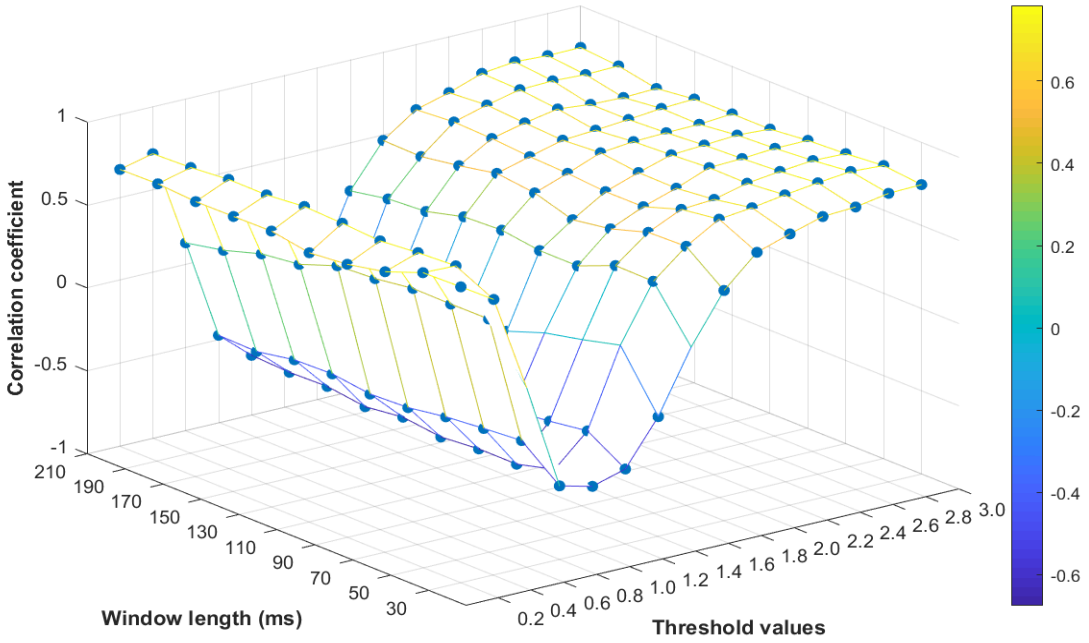


Figure 5-7: Correlation coefficient between STOI and LMPSC-SRBIM calculated using different combinations of window length and threshold values in picking peaks. Dots signify points of statistical significance ($p < 0.001$). Threshold had a larger effect on correlation than window length

Within threshold of 1.0, MFCC-SRBIM showed the greatest negative correlation of -0.689 ±0.005 at 90 ms, the same window length chosen through ESTOI correlations. For LMPSC-SRBIM, the largest negative correlations were observed at 150 ms window length with -0.676 ±0.005, and at 110 ms window length with a correlation of -0.675 ±0.004. Both values were within their respective error ranges, calculated through the leave-one-out approach implemented on the thirty speech files, suggesting that the difference in correlation magnitudes between these two window lengths were statistically insignificant.

Correlation of MFCC-SRBIM to STOI showed the most negative correlation at threshold 1.0 and window length 90 ms, which was in complete agreement with the choice suggested with ESTOI correlations. For STOI correlation with LMPSC-SRBIM, the greatest negative correlations at threshold 1.0 were observed at window lengths 150 ms and 110 ms, suggesting that the latter was the optimal window length for LMPSC-SRBIM, as this choice derived through analysis of ESTOI correlations. The correlation behaviour of SNR loss to SRBIM in the next section was used to finalize the optimized values of the two parameters to be used for the peak picking algorithms for the later simulations.

5.1.3 Correlation with SNR Loss for Parameter Optimization

Thus far, ESTOI and STOI correlations have suggested that MFCC-SRBIM should be calculated using threshold of 1.0 and window length of 90 ms. Similar correlation analysis with the LMPSC-SRBIM suggested that the optimized threshold parameter was also 1.0, with optimized window length to be either 110 ms or 150 ms. Window length of 110 ms seems to be the most logical candidate as it showed the most negative correlations with both ESTOI and STOI. SNR loss correlation with SRBIM was used as the final judge on the choice of optimized parameters.

SNR loss was a metric that was expected to have positive correlation with SRBIM. As its name suggests, SNR loss reflects the degree of loss in intelligibility in the target speech as compared to its clean reference. Its value ranges from [0, 1], with 0 indicating no loss in intelligibility. As such, an increase in SNR loss means a decrease in speech intelligibility, which was the same behaviour as that expected with SRBIM.

Indeed, positive correlations between SRBIM and SNR loss were observed at threshold 1.0, as shown in Figure 5-8 for MFCC-SRBIM, and Figure 5-9 for LMPSC-SRBIM. Similar to ESTOI and STOI, while threshold had a large effect on both magnitude and directionality of the observed correlation, window length had a much smaller effect. Threshold 1.0 had the largest positive correlation, suggesting 1.0 to be the optimized threshold value to use when calculating SRBIM for later simulations.

With regards to window length, MFCC-SRBIM had the greatest positive correlation of 0.537 ±0.005 at window length of 90 ms. This was in complete agreement with the correlations observed with ESTOI and STOI. For LMPSC-SRBIM, the greatest positive correlation of 0.550 ±0.005 was observed with the window length of 110 ms. Thus, 110 ms window length was chosen to be the optimal window length for calculating LMPSC-SRBIM as it produced the most negative correlation with all three existing intelligibility metrics. The complete table of correlation coefficients can be seen in Appendix C for MFCC-SRBIM, and Appendix D for the LMPSC-SRBIM.

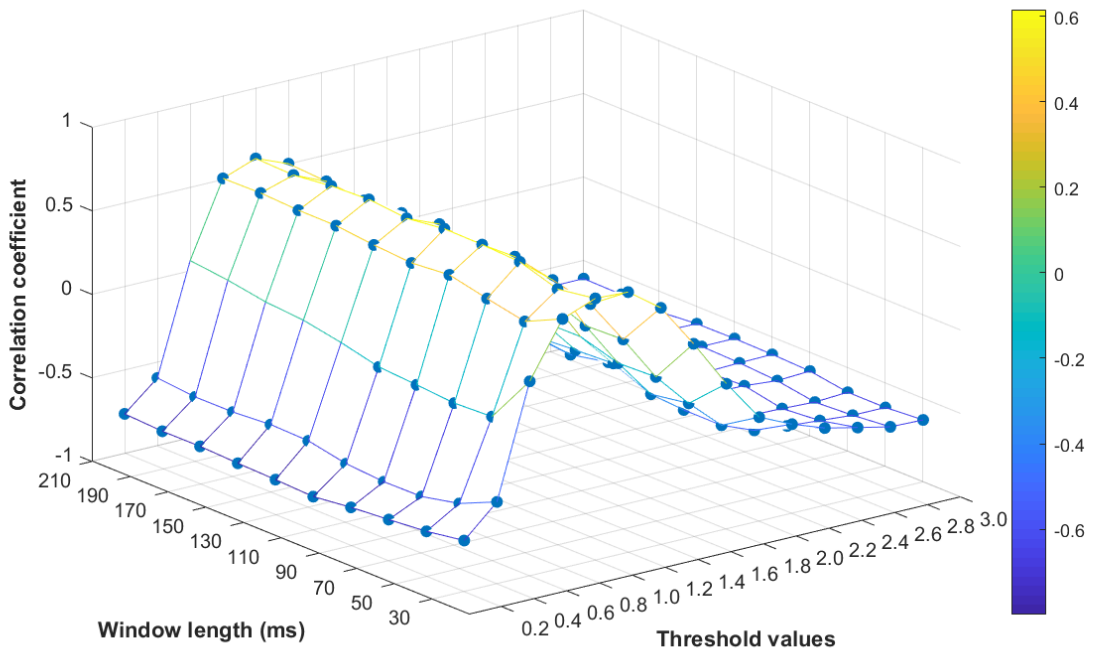


Figure 5-8: Correlation coefficient between SNR loss and MFCC-SRBIM calculated using different combinations of window length and threshold values in picking peaks. Dots signify points of statistical significance ($p < 0.001$). Threshold had a larger effect on correlation than window length

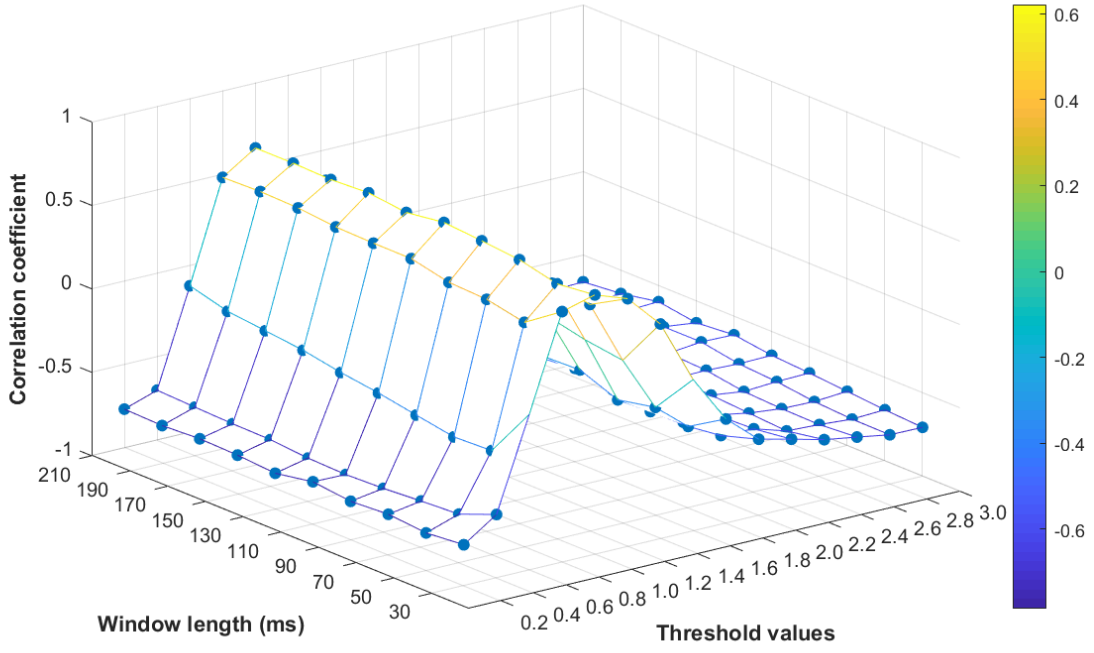


Figure 5-9: Correlation coefficient between SNR loss and LMPSC-SRBIM calculated using different combinations of window length and threshold values in picking peaks. Dots signify points of statistical significance ($p < 0.001$). Threshold had a larger effect on correlation than window length

Peak picking threshold parameter was optimized to 1.0, and the peak picking window length parameter was optimized to 90 ms for MFCC-SRBIM, and 110 ms for LMPSC-SRBIM. This decision was based predominantly on the correlation magnitudes and directionalities between SRBIM and the three existing metrics of ESTOI, STOI and SNR loss. Some theory and expectations of the desired behaviour of SRBIM were also considered when deciding on directionality. The next section evaluated SRBIM performance by comparing them to the three existing metrics via correlations using noisy speech stimuli.

5.2 Correlation of Metrics in Evaluating Noisy Speech

After parameters were optimized for MFCC-SRBIM and LMPSC-SRBIM to yield the strongest negative correlation with ESTOI and STOI, and the strongest positive correlation with SNR loss, the behaviours of the metrics were compared in more detail. Namely, correlations and scatter between the metrics were analyzed using noisy, but otherwise unprocessed speech, consisting of thirty-two noise variants of each of the thirty speech stimuli, and its clean version.

ESTOI, STOI, SNR loss, MFCC-SRBIM and LMPSC-SRBIM were calculated for all the speech stimuli, consisting of its clean version and noisy versions, which varied across eight types of noise presented at four levels of loudness. The correlation between any two metrics would reveal the strength and directionality of the relationship, and a scatter plot would further uncover the general trend in the behaviour of the metrics. Since correlations are better calculated through a collection of samples that varied in its range of speech intelligibility, the four noise levels, which were expected to be a large contributor to intelligibility, were kept intact when testing the noise types. In other words, comparison of the metrics was carried out nine times. Eight comparisons were for the eight types of noise, each containing all four levels of noise and the clean version of the speech stimuli. The last comparison was the aggregate data containing all types and levels of noise, and its clean version.

The intent of the comparison of the metric behaviours through correlation was to evaluate SRBIM performance in measuring speech intelligibility indirectly through the existing measures. A high correlation value would infer that there was a high degree of similarity in the behaviour of the compared metrics. Previous studies have already found that the existing metrics have relatively high correlation with human ratings of speech intelligibility or quality in certain situations. As such, the degree of correlation between SRBIM and existing measures gives an indication of how well the new metrics may be able to track intelligibility.

However, this indirect comparison also meant that results should be analyzed with care. A high correlation does not necessarily mean high metric performance, and a lower correlation does not necessarily mean low performance. Correlation comparisons reflect the degree of similarity between two metrics, which does not directly equate to metric performance in measuring intelligibility. As such, correlations will be visualized through scatter plots for further analysis of the effects of noise levels or speech enhancement on the outcome of the metrics. This was one of the advantages of incremental testing data, where comparisons across slightly different stimuli were possible. Additionally, correlations between existing measures were also presented to give the level of similarity between existing and accepted measures.

The calculated correlations between the metric pairs using noisy stimuli are summarized in Table 5-3 for the eight noise types and Table 5-4 for the aggregate data, where all correlations were found to have statistical significance ($p < 0.001$). In general, the directionality of the relationship was easily judged by whether the correlation was a positive or a negative value. However, the strength of the relationship determined from the magnitude can be challenging to interpret, as it can be affected by sample size, measurement error and the type of variables evaluated (Portney & Watkins, 2015). One book provided a general guideline of correlations between 0 and 0.25 to have little to no relationship, 0.25 to 0.5 to be considered a fair relationship, 0.50 to 0.75 to be in the range of a moderate to good relationship, and anything above 0.75 to constitute as good to excellent (Portney & Watkins, 2015). However, it also stated to exercise caution in interpreting correlation magnitude, suggesting these ranges as a starting point (Portney & Watkins, 2015).

Using these values as a guide, some general observations were made from Table 5-3 and Table 5-4. The most prevalent observation, which was a strong indication of SRBIM performance in measuring noisy speech intelligibility, was the strength of the correlation between SRBIM and existing metrics in comparison to the correlations within existing metrics. In general, a lower correlation was observed between SRBIM and existing measures than the correlations between ESTOI, STOI and SNR loss. Even with the noise wise analysis, which reported higher magnitudes than the aggregate data, correlations between SRBIM and existing metrics ranged in magnitude from the higher 0.60's to mid 0.80's. In contrast, ESTOI, STOI and SNR loss ranged in correlation magnitude from the mid 0.70's to mid 0.90's. There were noise

types, where the reported correlation magnitude between SRBIM and the existing metrics were within range of some of the correlations within existing metrics.

Table 5-3: Correlation coefficients between the five metrics for the eight types of noise using unprocessed noisy speech

		Airport	Babble	Car	Exhibition	Restaurant	Station	Street	Train
MFCC-SRBIM vs	ESTOI	-0.778 ±0.006	-0.779 ±0.006	-0.846 ±0.004	-0.785 ±0.005	-0.716 ±0.008	-0.817 ±0.005	-0.777 ±0.005	-0.772 ±0.006
	STOI	-0.766 ±0.006	-0.773 ±0.005	-0.847 ±0.005	-0.779 ±0.005	-0.719 ±0.008	-0.802 ±0.007	-0.766 ±0.006	-0.759 ±0.005
	SNR Loss	0.728 ±0.007	0.702 ±0.007	0.761 ±0.005	0.688 ±0.005	0.671 ±0.008	0.743 ±0.006	0.728 ±0.006	0.703 ±0.005
	LMPSC-SRBIM	0.842 ±0.005	0.850 ±0.005	0.896 ±0.003	0.834 ±0.006	0.816 ±0.008	0.893 ±0.004	0.889 ±0.004	0.851 ±0.005
LMPSC-SRBIM vs	ESTOI	-0.779 ±0.007	-0.811 ±0.005	-0.839 ±0.004	-0.771 ±0.006	-0.757 ±0.007	-0.798 ±0.007	-0.759 ±0.006	-0.764 ±0.006
	STOI	-0.756 ±0.005	-0.783 ±0.004	-0.817 ±0.005	-0.748 ±0.007	-0.744 ±0.006	-0.789 ±0.007	-0.756 ±0.006	-0.762 ±0.005
	SNR Loss	0.735 ±0.008	0.726 ±0.006	0.786 ±0.005	0.697 ±0.005	0.714 ±0.009	0.731 ±0.008	0.725 ±0.007	0.705 ±0.005
ESTOI vs	STOI	0.962 ±0.001	0.964 ±0.001	0.963 ±0.001	0.963 ±0.001	0.957 ±0.001	0.964 ±0.001	0.962 ±0.001	0.961 ±0.001
	SNR Loss	-0.905 ±0.001	-0.903 ±0.001	-0.911 ±0.001	-0.866 ±0.002	-0.892 ±0.001	-0.903 ±0.001	-0.902 ±0.001	-0.890 ±0.001
STOI vs	SNR Loss	-0.823 ±0.002	-0.820 ±0.002	-0.825 ±0.002	-0.774 ±0.003	-0.803 ±0.002	-0.823 ±0.002	-0.825 ±0.002	-0.806 ±0.002

Table 5-4: Correlation coefficients between the five metrics for the aggregate data using unprocessed noisy speech

MFCC-SRBIM vs	ESTOI	-0.694 ±0.005
	STOI	-0.689 ±0.005
	SNR Loss	0.613 ±0.008
	LMPSC-SRBIM	0.821 ±0.004
LMPSC-SRBIM vs	ESTOI	-0.687 ±0.005
	STOI	-0.675 ±0.004
	SNR Loss	0.621 ±0.008
ESTOI vs	STOI	0.951 ±0.002
	SNR Loss	-0.854 ±0.002
STOI vs	SNR Loss	-0.781 ±0.004

Specifically, for exhibition and car type noise, correlations reported between MFCC-SRBIM and STOI was higher than the correlation between STOI and SNR loss. As seen in Table 5-3, STOI and SNR loss had correlations at -0.774 ±0.003 for exhibition noise and -0.825 ±0.002 for car noise. These values were lower or equal in magnitude than the correlation between STOI and MFCC-SRBIM rate at -0.779 ±0.005 for exhibition and -0.847 ±0.005 for car type noise. In these two noise types, the similarities in correlation magnitudes suggested that MFCC-SRBIM's ability to measure intelligibility was comparable to the differences reported within various existing metrics, such as SNR loss as compared with STOI.

However, a closer inspection of the scatter plots of the metrics revealed a shortcoming of relying on only the correlation coefficients. While car type noise showed similar correlation magnitudes between STOI and SNR loss, and STOI and MFCC-SRBIM, the metric behaviours were quite different. Figure 5-10 shows a plot of SNR loss against STOI for the speech files containing car type noise, where SNR loss has a nonlinear relationship with STOI, with a large jump in SNR loss from 0 and 0.4. While STOI quantified the speech files with car noise at a SNR of 15 dB to have an intelligibility around 0.9, close to the clean speech intelligibility of 1, SNR loss detected this difference to be quite large. In fact, this gap in SNR loss resulted in all comparisons involving SNR loss to be nonlinear.

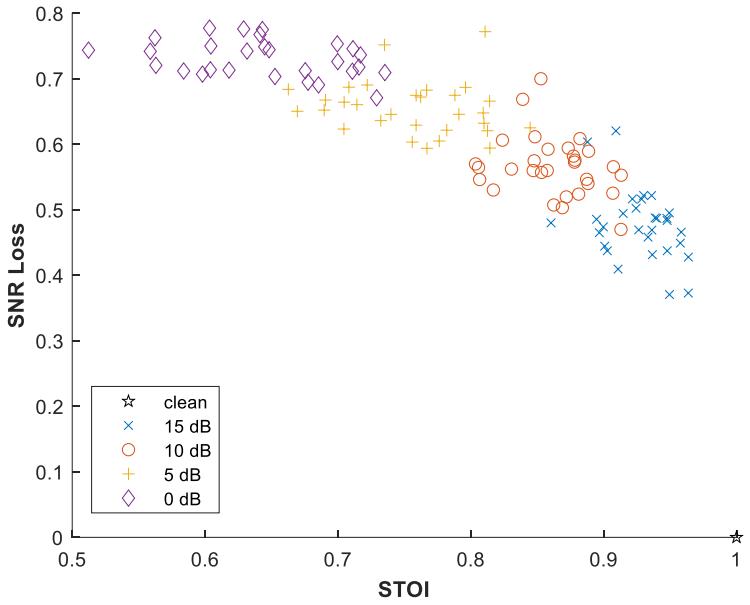


Figure 5-10: Scatter plot of SNR loss against STOI for noisy speech with car type noise

Conversely, the plot of MFCC-SRBIM to STOI in Figure 5-11 (a) shows that the plots were more scattered, but much of the speech stimuli presented at SNR 15 dB had a MFCC-SRBIM close to the clean value. In fact, Figure 5-11 (a) suggests that while there are variances in the outcomes, for an adequately large sample, MFCC-SRBIM may be a sufficient measure of speech intelligibility for simply noisy speech, as a roughly linear relationship was observed with STOI. Referring to the rough guidelines of the magnitude of the correlations and the strength of the relationships, correlations above 0.75, which were the majority of the correlations observed between SRBIM and ESTOI or STOI analyzed for each type of noise, were in the good to excellent range (Portney & Watkins, 2015).

A similar scatter plot was observed for LMPSC-SRBIM to STOI as shown in Figure 5-11 (b), as well as SRBIM and ESTOI, as seen in Figure 5-12. While each individual speech was prone to variance in its measured outcome, when analyzed as a collection of stimuli, both SRBIM showed similarity in performance to STOI, and ESTOI. SNR loss, on the other hand, showed a gap between the measured intelligibility of clean and SNR 15 dB speech stimuli in all its scatter plots, which may be due to the fact that SNR loss was originally designed for measuring intelligibility of noise reduced speech, tested in the next simulation. These results highlight the fact that SRBIM presented with a relatively strong linear relationship with ESTOI and STOI. On the other hand, SNR loss presented with an association that would have been better represented through nonlinear means, but resulted in correlation coefficients on par or better than the combinations observed between SRBIM and existing measures.

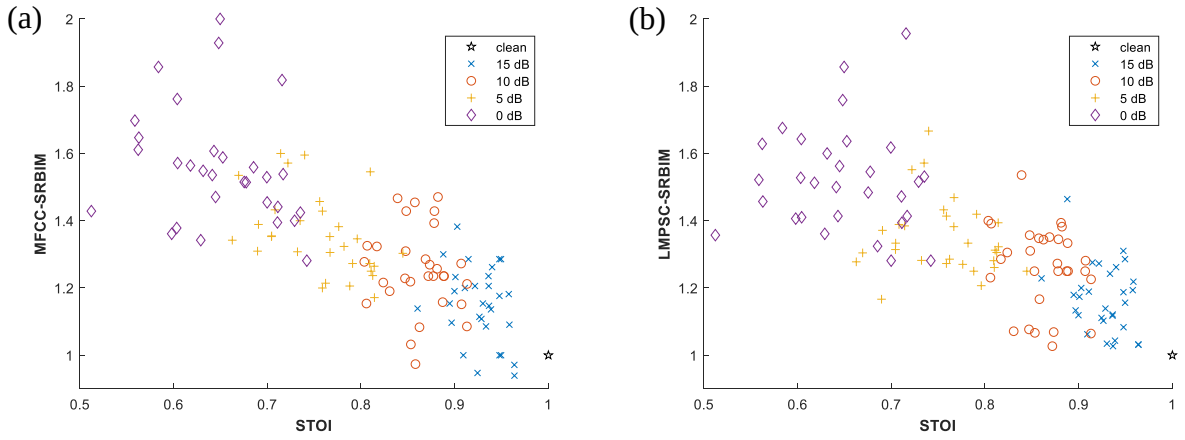


Figure 5-11: Scatter plot of (a) MFCC-SRBIM against STOI and (b) LMPSC-SRBIM against STOI for noisy speech with car type noise

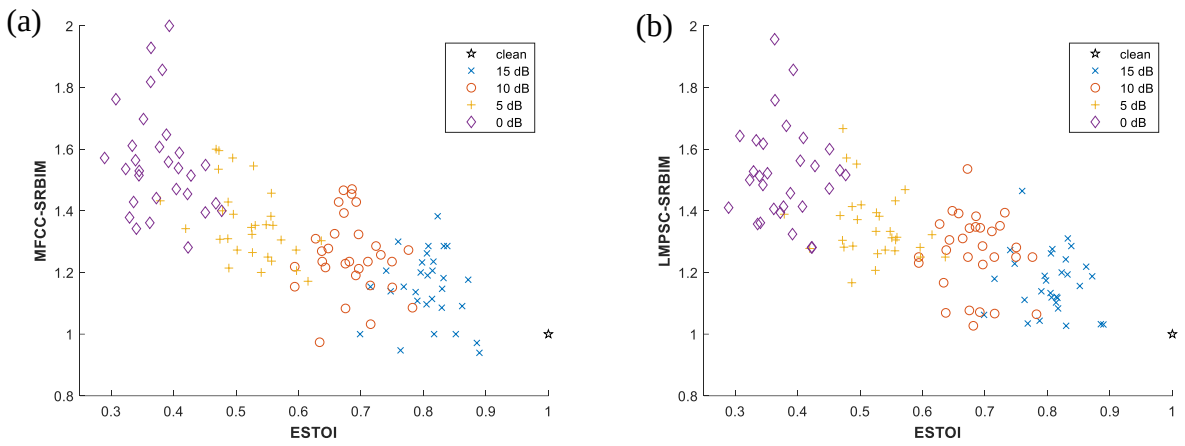


Figure 5-12: Scatter plot of (a) MFCC-SRBIM against ESTOI and (b) LMPSC-SRBIM against ESTOI for noisy speech with car type noise

Another observation was that MFCC-SRBIM and LMPSC-SRBIM had a lower correlation coefficient with the other metrics using aggregate data, compared to analysis that separated the noise types. Correlation between MFCC-SRBIM and ESTOI was -0.694 ±0.005 for the aggregate data, while the lowest correlation between the same metrics for the noise wise analysis was -0.716 ±0.008. Even with the errors, there is a difference of 0.022 ±0.013. Similar differences were observed between MFCC-SRBIM and STOI, and MFCC-SRBIM and SNR loss, where the correlation using aggregate data were 0.030 ±0.013 and 0.058 ±0.016 lower than the lowest noise wise correlations. Similar patterns were observed between LMPSC-SRBIM and the three existing metrics. Figure 5-13 shows the scatter plot of SRBIM to ESTOI with the aggregate data. The level of variance in the measured SRBIM has increased in comparison to the single noise analysis shown in Figure 5-12, which was the cause of the lower correlation in the aggregate data.

In comparison, the difference between the aggregate and noise wise correlations within the existing metrics was much smaller. For ESTOI and SNR loss, the difference was 0.012 ±0.004, and the difference for ESTOI and STOI was 0.006 ±0.003. The aggregate correlation between STOI and SNR loss was higher than the lowest noise type wise correlation observed.

From this observation, it can be inferred that the three existing metrics perform similarly in quantifying intelligibility of speech regardless of the noise type present. This similarity was lacking in the newly proposed SRBIM. Furthermore, this discrepancy was most likely caused by the differences in the method of quantifying speech intelligibility. Specifically, ESTOI, STOI and SNR loss all measured speech intelligibility by comparing the target speech against the

reference speech in the frequency domain. SRBIM, on the other hand, converted the signals into the frequency domain to find the cepstral and spectral coefficients, but ultimately compared the target and reference speeches in the time domain. As such, SRBIM most likely presented with separately linear relationship against the three existing metrics for each noise type, which was not along a common line of best fit when aggregated into one database. This resulted in a larger range of variance and lower correlation for the calculated SRBIM when considering the aggregate data.

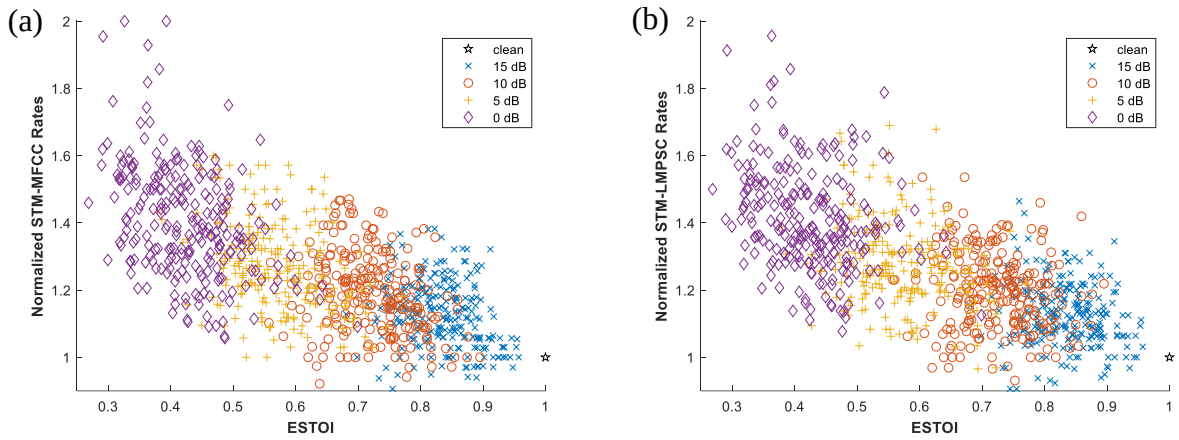


Figure 5-13: Scatter plot of (a) MFCC-SRBIM against ESTOI and (b) LMPSC-SRBIM against ESTOI for all noisy speech

Another surprising observation was the relatively high correlation of ESTOI with most of the metrics. While ESTOI and STOI correlation was expected to be high, as ESTOI was an extended version of STOI, ESTOI proved to be somewhat of a universal speech intelligibility score, showing some of the highest correlation magnitudes in all the tested comparisons.

In summary, correlation coefficients and the scatter plots of SRBIM to existing metrics, compared to the relationship between existing metrics, showed the potential of using SRBIM as a speech intelligibility metric when used in a collection of speech stimuli, and when noise was limited to one type. Analysis also revealed that SNR loss presented itself to have relatively high correlation magnitudes, especially with other existing measures, but the relationship was better represented by a nonlinear function, highlighting the disadvantage of using just correlation coefficients in analysis.

5.3 Summary of Simulation Results of Noisy Speech

This chapter presented the results of the optimization of the peak picking parameters needed when calculating MFCC-SRBIM and LMPSC-SRBIM, as well as the findings from the comparison of the metric behaviours for noisy, but otherwise unprocessed, speech. Threshold was observed to have a larger effect on the magnitude and direction of the correlation coefficients between SRBIM and the three existing metrics. Threshold 1.0 produced the desired behaviour of more speech boundaries detected with increased levels of noise, and was chosen as the optimized value. The window length used for subsequent calculations was 90 ms for MFCC-SRBIM and 110 ms for LMPSC-SRBIM. These were chosen for the largest magnitude in the correlation coefficients.

With analysis of the noisy, but otherwise unprocessed, speech, SRBIM was shown to have a relatively strong linear relationship with ESTOI and STOI, two of the existing speech intelligibility metrics. Analysis of the scatter plots revealed that SRBIM were linearly associated with ESTOI and STOI. However, scatter plots of the existing metric of SNR loss was shown to be nonlinearly related to ESTOI and STOI, undermining the assumption of the correlation analysis. These results indicate that the two newly proposed SRBIM have the potential in measuring the speech intelligibility of noisy speech.

6 Simulation Results of Noise Reduced Speech

The previous chapter presented the simulation results using speech that was noisy, but otherwise unprocessed. This chapter will describe the research findings associated with quantifying and comparing the intelligibility of speech stimuli that have been enhanced using several noise reduction algorithms. Analysis of the findings include correlation and scatter plots between pairs of metrics to investigate the similarity of the behaviour of the new metrics in relation to existing metrics. To add another dimension of analyzing the metric similarities, their agreement rates in ranking the efficiency of the noise reduction algorithms were also noted.

6.1 Correlation of Metrics in Evaluating Noise Reduced Speech

The five intelligibility metrics were measured for speech that had been noise reduced through three different noise reduction algorithms. These were the geometric approach to spectral subtraction, the probabilistic geometric approach to spectral subtraction, and the two-stage mel-warped Wiener filter approach. Similar to the simulations discussed in the previous chapter, the relative effectiveness of the metrics and noise reduction algorithm were analyzed through correlation coefficients as well as scatter plots. During analysis, the original noisy speech was also included to increase the stimuli range for better calculation of correlation coefficients. The inclusion of the original speech also helped clarify the similarities and differences in the metric behaviours for measuring simply noisy speech, as compared to metric behaviours in measuring noise reduced speech.

The correlation coefficients between the five metrics calculated for the noisy speech before and after it had been applied the geometric approach to spectral subtraction are summarized in Table 6-1. Many of the general observations from the first simulation described in Chapter 5 were seen again in this simulation. The correlations between SRBIM (Speech Rate Based Intelligibility Metric) and existing measures were lower than the correlations observed within existing measures. The correlations calculated for each noise type was generally higher than the correlation reported for the aggregate data. However, there was one difference from the results before.

As compared to correlations calculated using simply noisy speech in Chapter 5, the inclusion of the speech that had been passed through the noise reduction algorithms of the geometric approach to spectral subtraction, decreased the calculated correlations for

combinations of SRBIM to existing measures. In comparison, the decrease in the correlations between existing measures were less pronounced. This suggested that SRBIM significantly differed in how it measured speech intelligibility of noise reduced speech in comparison to simply noisy speech.

The difference in the SRBIM behaviours before and after noise reduction through the geometric approach to spectral subtraction can be visualized through Figure 6-1. The original noisy speech is plotted by the small markers, while the noise reduced speech is plotted by the large markers. In general, the noise reduced SRBIM were clustered in a lower range than the SRBIM calculated for noisy speech, which would be indicative of improved intelligibility through the geometric approach to spectral subtraction. However, from the general locations of the noise reduced plots, the same shift in ESTOI (Extended Short-Time Objective Intelligibility) measures were not observed after noise reduction, resulting in a lower correlation coefficient with the greater variance of the samples.

Table 6-1: Correlation coefficients between the five metrics for the eight types of noise using noisy speech before and after noise reduction through the geometric approach to spectral subtraction

		Airport	Babble	Car	Exhibition	Restaurant	Station	Street	Train
MFCC-SRBIM vs	ESTOI	-0.664 ±0.005	-0.670 ±0.006	-0.735 ±0.005	-0.641 ±0.007	-0.593 ±0.008	-0.711 ±0.005	-0.654 ±0.006	-0.630 ±0.006
	STOI	-0.638 ±0.006	-0.652 ±0.005	-0.719 ±0.005	-0.605 ±0.006	-0.567 ±0.007	-0.686 ±0.006	-0.636 ±0.007	-0.592 ±0.006
	SNR Loss	0.601 ±0.008	0.594 ±0.008	0.661 ±0.005	0.571 ±0.005	0.534 ±0.008	0.642 ±0.005	0.612 ±0.006	0.578 ±0.006
	LMPSC-SRBIM	0.842 ±0.004	0.856 ±0.004	0.886 ±0.003	0.839 ±0.005	0.816 ±0.005	0.884 ±0.003	0.880 ±0.003	0.839 ±0.005
LMPSC-SRBIM vs	ESTOI	-0.691 ±0.005	-0.718 ±0.005	-0.733 ±0.005	-0.639 ±0.005	-0.632 ±0.007	-0.729 ±0.005	-0.648 ±0.006	-0.652 ±0.006
	STOI	-0.656 ±0.004	-0.689 ±0.004	-0.701 ±0.005	-0.584 ±0.005	-0.611 ±0.006	-0.712 ±0.005	-0.635 ±0.006	-0.628 ±0.006
	SNR Loss	0.635 ±0.007	0.635 ±0.006	0.683 ±0.005	0.609 ±0.005	0.574 ±0.009	0.657 ±0.007	0.624 ±0.007	0.612 ±0.006
ESTOI vs	STOI	0.961 ±0.001	0.964 ±0.001	0.964 ±0.001	0.962 ±0.001	0.958 ±0.001	0.964 ±0.001	0.962 ±0.001	0.963 ±0.001
	SNR Loss	-0.884 ±0.001	-0.882 ±0.001	-0.889 ±0.001	-0.848 ±0.002	-0.871 ±0.001	-0.886 ±0.001	-0.882 ±0.001	-0.871 ±0.001
STOI vs	SNR Loss	-0.810 ±0.002	-0.805 ±0.002	-0.815 ±0.002	-0.759 ±0.003	-0.791 ±0.002	-0.812 ±0.002	-0.810 ±0.002	-0.796 ±0.002

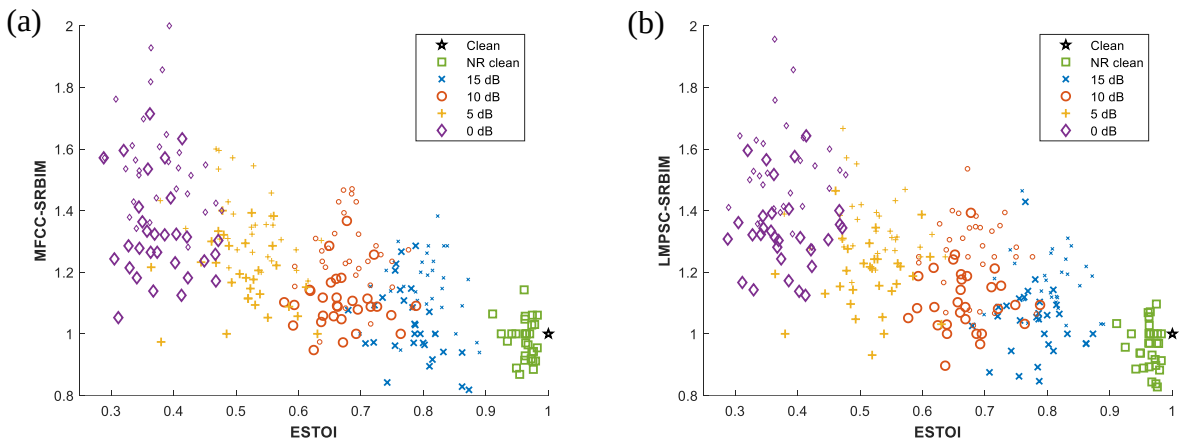


Figure 6-1: Scatter plot of (a) MFCC-SRBIM against ESTOI and (b) LMPSC-SRBIM against ESTOI for noisy speech (small markers) and noise reduced (NR) speech via geometric approach to spectral subtraction (large markers) with car type noise. Both SRBIM are linearly related to ESTOI, but greater variance is observed with the inclusion of noise reduced speech samples

In comparison, there were not much of an increase in the variance of the location of the speech stimuli in the scatter plot between two highly similar speech intelligibility measures. For example, in Figure 6-2, the noisy (small markers) and noise reduced (large markers) speech stimuli show a great level of overlap between ESTOI and STOI (Short-Time Objective Intelligibility). While some difference in the metric behaviours were observed for noise level of SNR 15 dB, most other speech samples in the other noise levels for before and after noise reduction via the geometric approach were located on the same linear function of best fit. This resulted in comparable correlation coefficients before and after noise reduction for the existing measures.

The other methods of noise reduction provided similar observations to the geometric approach to spectral subtraction. Namely, there was a larger decrease in correlation magnitude comparing SRBIM to existing measures as opposed to pairs within the three existing measures using speech stimuli consisting of noisy speech as well as noise reduced speech, regardless of the type of noise reduction algorithm used. The correlation coefficients for the noise type wise analysis of the probabilistic geometric approach to spectral subtraction are summarized in Table 6-2, and the correlations of the noise type wise analysis of the two-stage mel-warped Wiener filter approach is summarized in Table 6-3. While the magnitude of the decrease in correlation coefficients was more pronounced with SRBIM, the general pattern of decreasing correlations depending on the type of noise reduction implemented was consistent between most metric pairs.

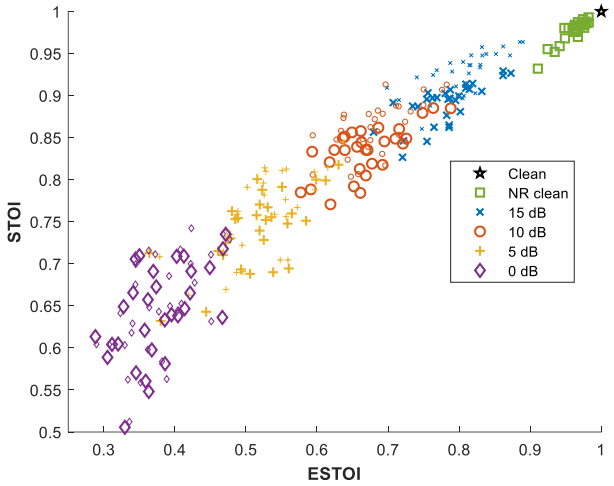


Figure 6-2: Scatter plot of STOI against ESTOI for noisy speech (small markers) and noise reduced (NR) speech via geometric approach to spectral subtraction (large markers) with car type noise. For two similarly behaving metrics, the plots before and after noise reduction aggregate on the same linear relationship

Table 6-2: Correlation coefficients between the five metrics for the eight types of noise using noisy speech before and after noise reduction through the probabilistic geometric approach to spectral subtraction

		Airport	Babble	Car	Exhibition	Restaurant	Station	Street	Train
MFCC-SRBIM vs	ESTOI	-0.704 ±0.006	-0.733 ±0.006	-0.769 ±0.004	-0.724 ±0.005	-0.692 ±0.007	-0.742 ±0.005	-0.715 ±0.005	-0.699 ±0.007
	STOI	-0.696 ±0.006	-0.731 ±0.005	-0.771 ±0.004	-0.718 ±0.005	-0.686 ±0.008	-0.729 ±0.006	-0.713 ±0.005	-0.686 ±0.006
	SNR Loss	0.681 ±0.007	0.671 ±0.007	0.733 ±0.004	0.664 ±0.005	0.661 ±0.008	0.716 ±0.006	0.698 ±0.005	0.663 ±0.005
	LMPSC-SRBIM	0.817 ±0.006	0.852 ±0.004	0.883 ±0.003	0.817 ±0.006	0.814 ±0.007	0.879 ±0.003	0.868 ±0.004	0.829 ±0.005
LMPSC-SRBIM vs	ESTOI	-0.727 ±0.006	-0.753 ±0.005	-0.769 ±0.005	-0.693 ±0.006	-0.726 ±0.007	-0.749 ±0.005	-0.688 ±0.007	-0.723 ±0.005
	STOI	-0.701 ±0.005	-0.736 ±0.004	-0.762 ±0.005	-0.666 ±0.007	-0.711 ±0.007	-0.738 ±0.006	-0.689 ±0.007	-0.709 ±0.005
	SNR Loss	0.707 ±0.007	0.701 ±0.007	0.753 ±0.005	0.673 ±0.005	0.691 ±0.008	0.725 ±0.006	0.692 ±0.007	0.694 ±0.005
ESTOI vs	STOI	0.963 ±0.001	0.966 ±0.001	0.966 ±0.001	0.963 ±0.001	0.958 ±0.001	0.966 ±0.001	0.965 ±0.001	0.963 ±0.001
	SNR Loss	-0.9041 ±0.0008	-0.905 ±0.001	-0.9107 ±0.0009	-0.876 ±0.002	-0.892 ±0.001	-0.9026 ±0.0008	-0.901 ±0.001	-0.885 ±0.001
STOI vs	SNR Loss	-0.828 ±0.002	-0.827 ±0.002	-0.835 ±0.002	-0.786 ±0.003	-0.806 ±0.002	-0.829 ±0.002	-0.829 ±0.002	-0.806 ±0.002

Table 6-3: Correlation coefficients between the five metrics for the eight types of noise using noisy speech before and after noise reduction through the two-stage mel-warped Wiener filter approach

		Airport	Babble	Car	Exhibition	Restaurant	Station	Street	Train
MFCC-SRBIM vs	ESTOI	-0.664 ±0.007	-0.665 ±0.006	-0.701 ±0.004	-0.641 ±0.007	-0.620 ±0.007	-0.680 ±0.005	-0.649 ±0.006	-0.605 ±0.006
	STOI	-0.645 ±0.007	-0.660 ±0.005	-0.694 ±0.005	-0.642 ±0.005	-0.600 ±0.007	-0.664 ±0.006	-0.643 ±0.006	-0.565 ±0.006
	SNR Loss	0.592 ±0.009	0.588 ±0.008	0.646 ±0.006	0.588 ±0.006	0.553 ±0.009	0.634 ±0.006	0.605 ±0.006	0.580 ±0.006
	LMPSC-SRBIM	0.839 ±0.004	0.857 ±0.004	0.891 ±0.003	0.839 ±0.004	0.824 ±0.006	0.880 ±0.003	0.873 ±0.003	0.854 ±0.004
LMPSC-SRBIM vs	ESTOI	-0.649 ±0.005	-0.690 ±0.006	-0.708 ±0.004	-0.615 ±0.006	-0.619 ±0.008	-0.663 ±0.005	-0.619 ±0.007	-0.592 ±0.006
	STOI	-0.619 ±0.005	-0.661 ±0.005	-0.687 ±0.006	-0.594 ±0.006	-0.592 ±0.008	-0.653 ±0.006	-0.620 ±0.007	-0.563 ±0.006
	SNR Loss	0.588 ±0.008	0.606 ±0.008	0.664 ±0.006	0.595 ±0.006	0.55 ±0.01	0.626 ±0.007	0.597 ±0.007	0.580 ±0.006
ESTOI vs	STOI	0.962 ±0.001	0.964 ±0.001	0.966 ±0.001	0.962 ±0.001	0.958 ±0.001	0.964 ±0.001	0.963 ±0.001	0.957 ±0.001
	SNR Loss	-0.841 ±0.002	-0.844 ±0.002	-0.849 ±0.002	-0.815 ±0.002	-0.828 ±0.002	-0.839 ±0.002	-0.832 ±0.002	-0.826 ±0.002
STOI vs	SNR Loss	-0.778 ±0.002	-0.779 ±0.003	-0.791 ±0.003	-0.743 ±0.003	-0.755 ±0.002	-0.785 ±0.003	-0.773 ±0.002	-0.749 ±0.003

Figure 6-3 provides a visual representation of the magnitudes of correlation for various metric pairs with speech stimuli before and after noise reduction through the three approaches. Correlation magnitude between ESTOI and STOI remained high regardless of the presented stimuli, but correlations between ESTOI and SNR (signal to noise ratio) loss, and between STOI and SNR loss varied with the type of noise reduction implemented. The pattern of variation was similar with metric pairs between SRBIM and existing measures.

As a general trend, the geometric approach to spectral subtraction produced correlations magnitudes lower than that observed for simply noisy speech. The Wiener filter design had the lowest correlations for SRBIM compared to existing measures, as well as ESTOI compared to SNR loss and STOI compared with SNR loss. However, there was a slight difference observed for correlation differences between the original noisy speech and those applied the probabilistic geometric approach to spectral subtraction. While SRBIM to existing measures for the probabilistic geometric approach showed a consistent decrease in correlation magnitudes, ESTOI and STOI comparisons to SNR loss showed the same or slight increase in these correlations.

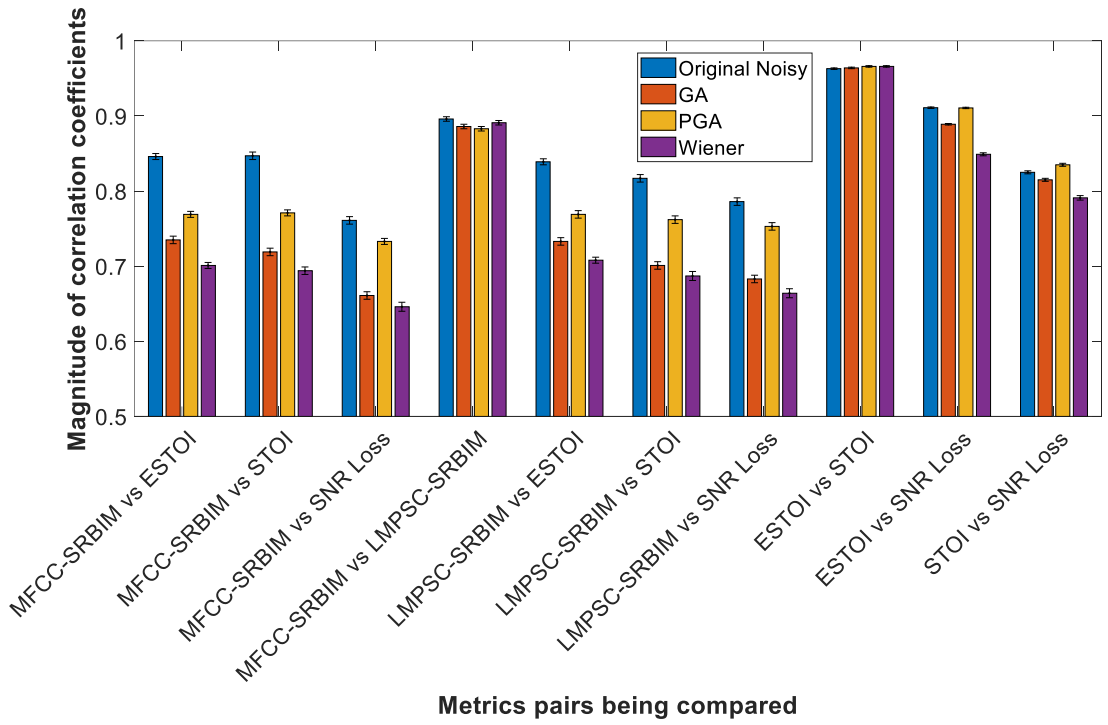


Figure 6-3: Magnitude of the correlations between various combinations of the five metrics for the original noisy speech (blue), including speech before and after geometric approach (GA) to spectral subtraction (red), before and after probabilistic geometric approach (PGA) to spectral subtraction (yellow), and two-stage mel-warped Wiener filter approach (purple). The correlations were calculated using car type noisy speech. The relative correlation magnitudes between the three noise reduction algorithms were similar between SRBIM and existing metrics were similar to comparison of ESTOI or STOI to SNR loss.

Regardless of the difference in the relation between the noisy speech correlations to the noise reduced speech correlations, the relative correlation differences between the three noise reduction algorithms were consistent through most of the metric pairs. Specifically, the Wiener approach produced the weakest correlation, and the probabilistic geometric approach observed the strongest correlation.

A possible explanation to this phenomenon of a larger decrease in correlation magnitudes after noise reduction with SRBIM to existing metric pairs may be, again, due to the differences in the domain of comparison for the metrics. Previously, it was mentioned that all three existing speech intelligibility metrics calculated intelligibility of a target speech in reference to the clean speech through comparisons in the frequency domain. The newly proposed SRBIM compared the two speech signals in the time domain. The difference in the domain of comparison may result in much larger differences in the pattern of output intelligibilities after noise reduction, as many of these algorithms reduce noise in the frequency domain. As support for this explanation,

the comparison of MFCC-SRBIM to LMPSC-SRBIM in Figure 6-3 remained relatively constant regardless of the type of noise reduction implemented. The smaller general decrease in correlation magnitudes observed for ESTOI and STOI correlations to SNR loss may be an extension of this hypothesis.

However, the linearity in the relationship between SRBIM and ESTOI, as well as SRBIM and STOI is preserved. As observed in Figure 6-1, although a larger variance in the scatter plots are observed between SRBIM and ESTOI, the rough approximation of the relationship is still linear. This was also true for comparisons of SRBIM to STOI. These results indicate that when taken together as a collective set of data containing one type of noise, SRBIM still exhibits a relatively strong similarity to ESTOI and STOI, which in turn suggests that SRBIM has potential in measuring intelligibility of noise reduced speech signals.

Correlation coefficients and scatter plots have been useful in visualizing how the speech intelligibility metrics differed in the way they quantified speech intelligibility of various stimuli before and after noise reduction. In general, correlations between SRBIM and existing measures experienced a larger decrease than correlations within existing measures with the inclusion of noise reduced speech stimuli. However, the general pattern of correlation magnitudes in the three noise reduction algorithms was consistent regardless of metric pairs compared, with the Wiener approach producing the weakest correlation, and the probabilistic geometric approach resulting in the strongest correlation. However, correlation coefficients only represent the similarity in the performances of two metrics. The next section momentarily assumes the validity of all five metrics in measuring speech intelligibility of noisy and noise reduced speech, to compare the performances of the three noise reduction algorithms.

6.2 Mock Experiment Comparing Noise Reduction Algorithms

This section used the five intelligibility metrics to compare the performances of the geometric approach to spectral subtraction, the probabilistic geometric approach to spectral subtraction, and the two-stage mel-warped Wiener filter approach to noise reduction. After the metrics were calculated for both noisy and noise reduced speech, their normalized differences were obtained for each of the three noise reduction algorithms implemented. This difference could be visualized as the relative degree to which speech intelligibility had been improved or compromised as a result of the application of a certain noise reduction algorithm. By comparing these relative improvements between the different noise reduction schemes through paired t-

tests, each metric could rank the effective performance of the three noise reduction algorithms, much like an actual experiment evaluating algorithm performance. The metrics were compared in their generated rankings, and general agreement within metric groups were observed. Unlike correlation coefficients, the noise levels could be separated for analysis. This allowed analysis using the aggregate data combining all noise types and levels, analysis separated by the eight noise types, analysis separated by the four noise levels, as well as analysis separated by the thirty-two noise conditions, differing in noise type and level. The section will start with the simpler analysis consisting of aggregate data.

The normalized improvements quantified by each of the five metrics in the mock experiment using aggregate data is shown in Figure 6-4. Looking at the relative ranking of the three noise reduction algorithm performances, all five metrics ranked the Wiener filter approach as the best approach. The mean normalized improvements were positive for all five metrics, implying relative improvements.

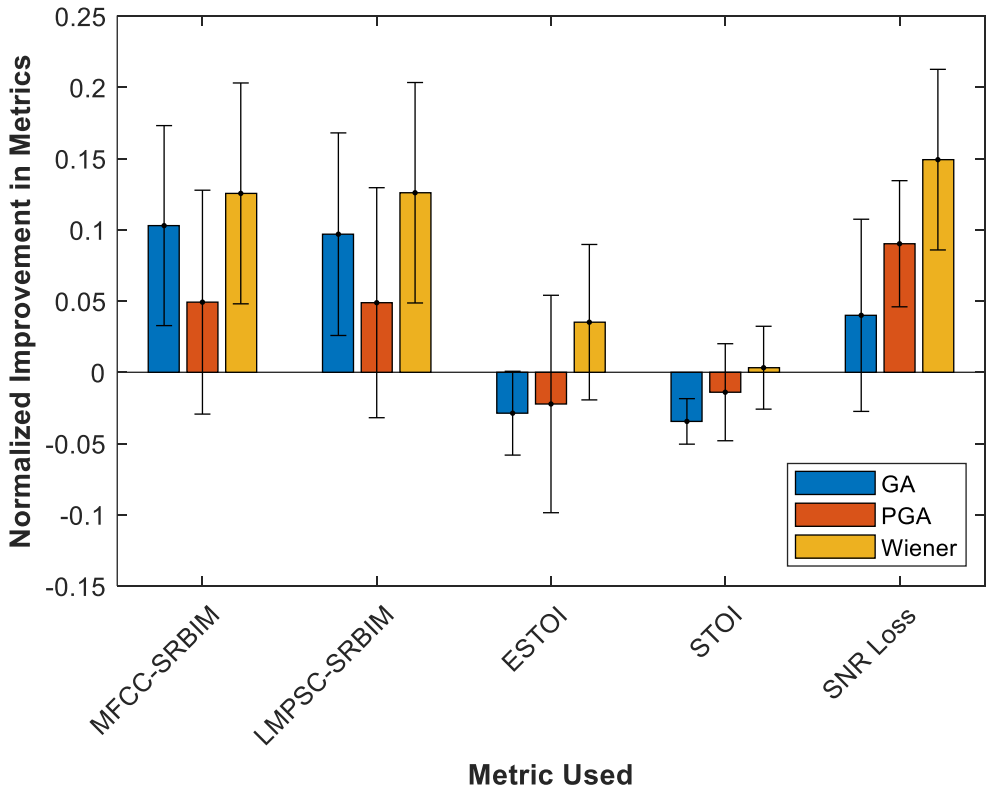


Figure 6-4: Mean normalized improvements for the five speech intelligibility metrics using aggregate data for three types of noise reduction algorithms of the geometric approach (GA) to spectral subtraction, probabilistic geometric approach (PGA) to spectral subtraction and two-stage mel-warped Wiener filter approach. Error bars encompass a range of two standard deviations from the mean value.

Table 6-4: Paired t-test comparison between two noise reduction algorithms (A and B) for aggregate data, and their results as judged by MFCC-SRBIM (red), LMPSC-SRBIM (yellow), ESTOI (blue), STOI (green) and SNR loss (grey)

A	B	A > B					A < B					A ≈ B				
Wiener	PGA															
Wiener	GA															
PGA	GA															

However, Figure 6-4 also shows that there was some disagreement in the ranking of the probabilistic geometric approach and the geometric approach to spectral subtraction. SRBIM ranked the geometric approach to be better than the probabilistic geometric approach, which was reversed in the three existing metrics. Additionally, paired t-tests using $p < 0.001$ was executed to test for statistical significances in these differences which supported the rankings. The results of these t-tests for the aggregate data are summarized in Table 6-4.

Each row in Table 6-4 represents a paired t-test comparison between two noise reduction algorithms, arbitrarily labeled “A” and “B”. The paired t-test was then used to determine whether the difference in speech improvements, and subsequently noise reduction algorithm performances, were statistically significant, resulting in a relative ranking of the algorithms for each metric. The differently coloured boxes represent each of the five metrics used, where red represents MFCC-SRBIM (Mel Frequency Cepstral Coefficient type Speech Rate Based Intelligibility Metric), yellow is LMPSC-SRBIM (Log Mel Power Spectral Coefficient type Speech Rate Based Intelligibility Metric), blue is ESTOI, green is STOI and grey is SNR loss.

For example, it can be observed from rows 1 and 2 in Table 6-4 that all five metrics thought that Wiener filter based noise reduction was superior to the other approaches. However, in row 3, while SRBIM judged the geometric approach to be better than the probabilistic geometric approach (A < B), ESTOI though the performances were not statistically different (A ≈ B), and STOI and SNR loss judged the probabilistic approach to be better (A > B). From these results, it can be inferred that the probabilistic geometric approach and the geometric approach had only slight differences that were not statistically significant in ESTOI, but significant in STOI and SNR loss. Additionally, even existing metrics could not come to a conclusive decision for the comparison of the performances of the probabilistic geometric approach and the geometric approach. In that regard, SRBIM had an equal level of agreement with existing metrics, showing that it was capable of judging Wiener design superiority over the other two approaches using aggregate data.

The database could also be separated in terms of noise level, rather than using aggregate data. Table 6-5 shows how each of the five metrics judged the paired t-tests of the normalized improvements caused by the three noise reduction algorithms. It can be observed that at SNR of 10 dB, 5 dB and 0 dB, the Wiener filter was judged to outperform the other approaches by at least two of the three existing metrics. SRBIM had perfect agreement with this judgment at all three noise levels. All five metrics also showed complete agreement in rating the geometric approach to have resulted in larger improvements in speech intelligibility than the probabilistic geometric approach at SNR 0 dB. The implication of the high level of agreement in SRBIM performance to existing metrics at high levels of noise was a gradual decrease in SRBIM agreement with decreasing noise.

At SNR 5 dB and 10 dB, the existing measures gradually migrated towards the reverse ordering between the geometric approach and the probabilistic geometric approach. On the other hand, SRBIM kept its rankings relatively constant. This migration resulted in SRBIM disagreeing with all three existing metrics at SNR 10 dB and 15 dB, in judging whether probabilistic geometric approach or the geometric approach was a superior noise reduction algorithm.

Table 6-5: Paired t-test comparison between two noise reduction algorithms (A and B) for noise level wise data, and their results as judged by MFCC-SRBIM (red), LMPSC-SRBIM (yellow), ESTOI (blue), STOI (green) and SNR loss (grey)

SNR	A	B	A > B					A < B					A ≈ B				
Clean	Wiener	PGA															
	Wiener	GA															
	PGA	GA															
15 dB	Wiener	PGA															
	Wiener	GA															
	PGA	GA															
10 dB	Wiener	PGA															
	Wiener	GA															
	PGA	GA															
5 dB	Wiener	PGA															
	Wiener	GA															
	PGA	GA															
0 dB	Wiener	PGA															
	Wiener	GA															
	PGA	GA															

Whether this reversal of the performances between the geometric approach and the probabilistic geometric approach is an actual phenomenon requires more tests. The subjective evaluation of the speech quality implemented with ten listeners in a previous paper showed that at SNR 10 dB, the probabilistic geometric approach contained less distortion than the geometric approach (Islam et al., 2018). However, this previous experiment did not test for intelligibility, nor did it test at noise levels other than SNR 10 dB. The subjective ratings were also compared by taking the mean of the ten listeners, and is difficult to state whether the differences are statistically significant.

Nonetheless, if the results of the higher speech quality of the probabilistic geometric approach over the geometric approach to spectral subtraction is to be considered to be reflective of its corresponding speech intelligibility, SRBIM seems to be misaligned with the correct ranking observed in the existing metrics. These results imply that SRBIM has better agreement with existing metric performances at higher levels of noise.

However, the findings from Table 6-5 indicate that even existing metrics disagree at lower levels of noise. There was another shift in the existing metrics in the ranking orders between Wiener and probabilistic geometric approach at lower noise levels. At SNR 10 dB, STOI judged the two approaches to have statistically insignificant differences, while at SNR 15 dB, STOI flipped the two rankings around and judged the probabilistic geometric approach to outperform the Wiener filter approach. This shift resulted in disagreement between the existing metrics in the ranking of the Wiener filter and probabilistic geometric approaches at SNR 15 dB. In other words, at lower noise levels, the existing metrics would decrease in their rate of agreement in the noise reduction algorithm performances, which was much like the disagreement observed with SRBIM previously.

With the clean speech, the ranking reflected how well the algorithms were in avoiding the introductions of distortions. However, there are high levels of disagreement within existing metrics when discerning the relative rankings. The comparison between Wiener filter and geometric approach showed some level of agreement where SNR loss judged the Wiener filter to be better performing, and both ESTOI and STOI, as well as SRBIM, judged the differences to be insignificant. Analysis of the noise reduced speech separated by the noise levels seem to indicate that all metrics are in high agreement with each other at high levels of noise. However, at low

levels of noise, both SRBIM and existing metrics start disagreeing. Apart from analyzing the speech stimuli according to the noise levels, it could be analyzed depending on the noise type.

Table 6-6 summarizes the analysis by separating the speech files into the eight noise types. In all noise types, at least two of the three existing speech intelligibility metrics agreed that the Wiener filter based noise reduction algorithm was superior to either the probabilistic geometric approach and the geometric approach to spectral subtraction. MFCC-SRBIM judged Wiener filter to be superior to the other approaches in airport, car and station noise, showing agreement in three of the eight noise types. LMPSC-SRBIM had higher agreement, supporting Wiener design superiority in airport, babble, car, exhibition, station and train type noise, showing agreement in six of the eight noise types. These results suggest that when analyzing noise reduction performance based on noise type, LMPSC-SRBIM may have a higher agreement rate with existing metrics in comparison to MFCC-SRBIM.

Table 6-6: Paired t-test comparison between two noise reduction algorithms (A and B) for noise type wise data, and their results as judged by MFCC-SRBIM (red), LMPSC-SRBIM (yellow), ESTOI (blue), STOI (green) and SNR loss (grey)

Noise Type	A	B	A > B					A < B					A ≈ B				
Airport	Wiener	PGA															
	Wiener	GA															
	PGA	GA															
Babble	Wiener	PGA															
	Wiener	GA															
	PGA	GA															
Car	Wiener	PGA															
	Wiener	GA															
	PGA	GA															
Exhibition	Wiener	PGA															
	Wiener	GA															
	PGA	GA															
Restaurant	Wiener	PGA															
	Wiener	GA															
	PGA	GA															
Station	Wiener	PGA															
	Wiener	GA															
	PGA	GA															
Street	Wiener	PGA															
	Wiener	GA															
	PGA	GA															
Train	Wiener	PGA															
	Wiener	GA															
	PGA	GA															

On the other hand, the ranking between the probabilistic geometric approach and the geometric approach to spectral subtraction was, once again, difficult to judge even among the existing metrics. Even when STOI and SNR loss agreed that the probabilistic approach was better, ESTOI, the universal speech intelligibility measure that was highly correlated with all other metrics, was consistently rating the performance differences to be statistically insignificant. At the same time, SRBIM were also adamant in ordering the geometric approach to be better performing than the probabilistic geometric approach, suggesting that there is a persisting difficulty in judging the ranking order between the two variants of the geometric approaches.

Analysis by separating the speech files into noise levels revealed that many of the existing and newly proposed metrics agreed that Wiener filter greatly outperformed the other approaches, especially at higher levels of noise. Noise type wise analysis saw the same agreement especially with LMPSC-SRBIM and most of the existing metrics. Apart from clean speech, where noise reduction algorithms were expected to introduce more distortions than improvements in speech intelligibility, SRBIM has shown high consistency across noise types and noise levels in its perceived ranking of the noise reduction algorithms. Namely, it judged Wiener filter and the geometric approach to both be better performing than the probabilistic geometric approach. The Wiener filter was commonly ranked to outperform the geometric approach, but this difference was sometimes calculated to be statistically insignificant. On the other hand, existing metrics generally judged Wiener filter to be the best noise reduction algorithm, but were more variable in their perceived rankings between the two variants of the geometric approach.

With the existing metrics, Wiener filter approach was judged to be better performing than either probabilistic geometric approach or geometric approach in many cases. However, the relative ranking of these latter two would vary depending on noise level and type under analysis, as well as the metric used to evaluate the improvements. Whether these nuances in differences in the ranking of the noise reduction algorithm performances are valid will require comparison of the metric performances against subjective human ratings.

When the analyses of the paired t-test comparisons were extended to each of the thirty-two noise conditions, it became apparent that many of the metrics because unable to judge the statistical difference based on the p value of 0.001, as more boxes were filled in the column labeled as $A \approx B$, representing statistically insignificant differences in the comparison. This was

because each noise condition contained only thirty speech stimuli, which required more obvious differences to be perceived for differences to be judged as statistically significant in Table 6-7.

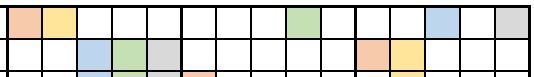
The MFCC-SRBIM and LMPSC-SRBIM were especially prone to be classifying differences as statistically insignificant with these smaller sample sizes, which was in agreement with the previous observation that a sufficient sample size was required to represent a roughly linear relationship between SRBIM and existing metrics of ESTOI and STOI.

6.3 Summary of Simulation Results of Noise Reduced Speech

In summary, the evaluation of the three noise reduction algorithms based on the five speech intelligibility measures revealed some findings. First, given that a large enough dataset was used, MFCC-SRBIM and LMPSC-SRBIM had a linear relationship with ESTOI and STOI, much like the simulation results using noisy speech from Chapter 5. Although the strength of the correlation had decreased in comparison to samples containing only noisy speech, analysis of the scatter plots show that the speech signals still aggregate to a roughly linear relationship. Additionally, through the mock experiment, SRBIM showed agreement with existing metrics, especially when using a large enough sample size, and especially at higher levels of noise. Even the existing metrics were prone to disagree in the algorithm performances at lower levels of noise. The best agreement in the existing metrics was observed as the superiority of the Wiener filter over the other two approaches. SRBIM had high agreement with ranking the Wiener filter as the best approach as well. As such, there was comparable level of agreement for all metrics, for ranking Wiener filter performance above the other approaches in many of the noise conditions, indicating that SRBIM can be useful in confirming the performances of algorithms that largely outperformed or underperformed the other approaches. The next chapter covers simulation findings based on speech that had been enhanced through various methods of rate modulation.

Table 6-7: Paired t-test comparison between two noise reduction algorithms (A and B) for data separated in noise type and noise level, and their results as judged by MFCC-SRBIM (red), LMPSC-SRBIM (yellow), ESTOI (blue), STOI (green) and SNR loss (grey)

Noise Type	SNR	A	B	A > B	A < B	A ≈ B
Airport	15 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			
	10 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			
	5 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			
	0 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			
Babble	15 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			
	10 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			
	5 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			
	0 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			
Car	15 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			
	10 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			
	5 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			
	0 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			
Exhibition	15 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			
	10 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			
	5 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			
	0 dB	Wiener	PGA			
		Wiener	GA			
		PGA	GA			

Restaurant	15 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	
	10 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	
	5 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	
	0 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	
Station	15 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	
	10 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	
	5 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	
	0 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	
Street	15 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	
	10 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	
	5 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	
	0 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	
Train	15 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	
	10 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	
	5 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	
	0 dB	Wiener	PGA	
		Wiener	GA	
		PGA	GA	

7 Simulation Results of Rate Modulated Speech

In the previous chapter, the linear relationship between SRBIM (Speech Rate Based Intelligibility Metrics) with ESTOI (Extended Short-Time Objective Intelligibility) or STOI (Short-Time Objective Intelligibility), and its agreement with existing metrics in rating Wiener filter noise reduction algorithm to be superior to other noise reduction algorithms, showed that SRBIM had potential to measure intelligibility of speech processed through noise reduction algorithms. This chapter presents the simulation findings corresponding to speech that had been processed through rate modulation algorithms. Analysis includes correlation and scatter plots between the two newly proposed SRBIM and three existing metrics, as well as the inclusion of some subjective observations made by me.

7.1 Correlation of Metrics in Evaluating Rate Modulated Speech

This simulation involved speech enhancement via rate modulating the speech, and comparison of the behaviour of the metrics before and after rate modulation. Both clean and noisy speech were rate modulated through four different algorithms, an existing uniform and non-uniform speech rate modulation algorithm, and two variants of the newly proposed SOLANU (SOLely Acoustic Non-Uniform) speech rate modulation algorithm. In the graphs, the existing uniform speech rate modulation was labeled as PSOLA-Uniform (Pitch Synchronous Overlap and Add), while the existing non-uniform speech rate modulation was labeled as PSOLA-Non-Uniform. For details on the two existing algorithms, please refer to Chapter 2. The two variants of SOLANU were called SOLANU-MFCC (Mel Frequency Cepstral Coefficients) and SOLANU-LMPSC (Log Mel Power Spectral Coefficients) depending on whether they used speech boundaries detected from the STM-MFCC (Spectral Transition Measures of Mel Frequency Cepstral Coefficients) or STM-LMPSC (Spectral Transition Measures of Log Mel Power Spectral Coefficients) signals.

Each algorithm underwent six degrees of speech stretch factors of 1.0, 1.2, 1.4, 1.6, 1.8 and 2.0. For PSOLA-Uniform, these stretch factors defined the constant value by which the resultant speech was time scaled. For the remaining non-uniform speech rate modulation algorithms, these were the maximum stretch factors used in the derivation of the time-varying stretch function. The minimum stretch value for non-uniform speech rate modulation algorithms were all fixed at 1.0. The six degrees of rate modulated variants of the speech were measured

through the five metrics, and the correlations between each metric pair was derived. The correlation coefficients as well as the scatter patterns were analyzed with respect to some personal observations made on the some of the speech files.

First, the effects of stretch factor on the magnitude of the correlations were investigated. Correlation magnitudes for the uniform speech rate modulation algorithm followed one of three patterns as a function of stretch factor. Figure 7-1 shows the correlation magnitudes using stimuli involving car type noise for each of the ten metric comparisons obtained for PSOLA-Uniform, the existing uniform speech rate modulation. For the two combinations of similar metrics, ESTOI compared to STOI (blue line at top of figure) and MFCC-SRBIM to LMPSC-SRBIM (purple line just below the blue line), the correlations were high, and remained high regardless of the stretch factor.

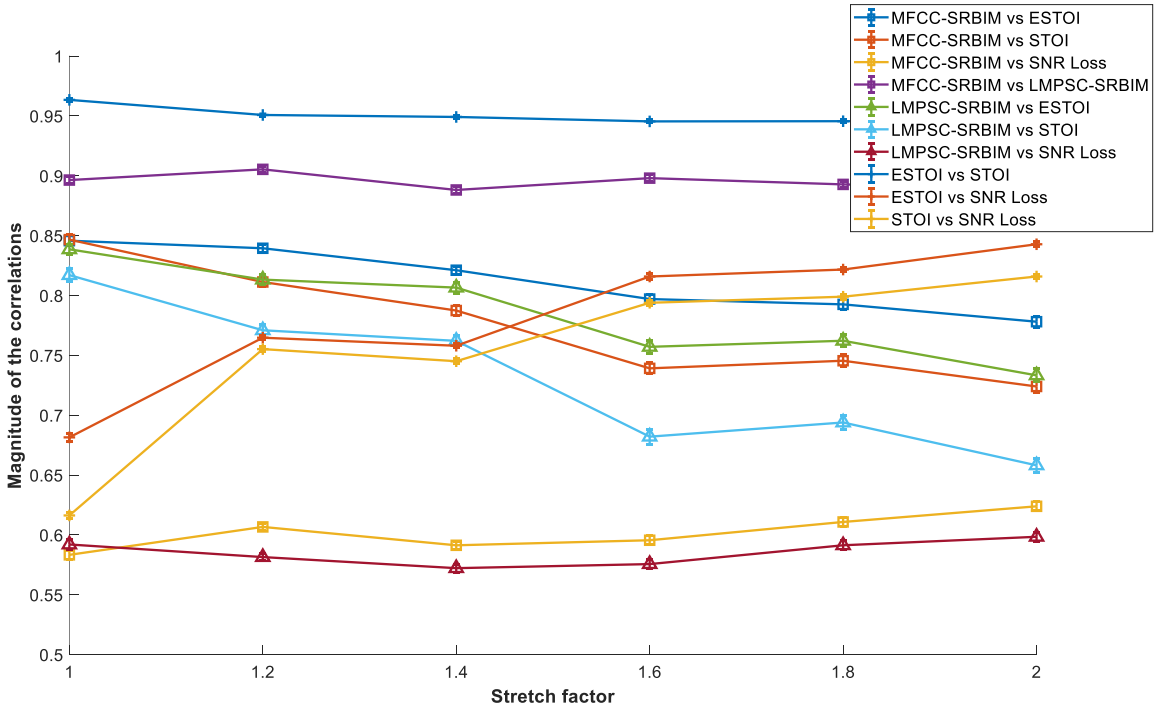


Figure 7-1: Magnitude of correlations between metric pairs for PSOLA-Uniform rate modulation for car type noise. Increase in stretch factor results in an increase in correlations between all metric pairs compared against SNR loss, and a decrease in correlations between SRBIM and either ESTOI or STOI.

On the other hand, SRBIM compared to either ESTOI (dark blue with square plots for MFCC-SRBIM vs ESTOI, green with triangles for LMPSC-SRBIM vs ESTOI) or STOI (red with square plots for MFCC-SRBIM vs STOI, light blue with triangles for LMPSC-SRBIM vs STOI) decreased with increasing stretch factor, indicating a greater variance in the behaviour of these two groups of metrics. The reverse behaviour was observed between any metric pair correlated against SNR (signal to noise ratio) loss, indicating an increase in the similarity of SNR loss towards the other metrics with increasing stretch factor used. The increase and decrease in correlation magnitudes can be explained by analyzing the scatter plots between two appropriately chosen metrics.

Figure 7-2 shows two scatter plots taken for the two extreme stretch factors of 1.0 (Figure 7-2 (a)) and 2.0 (Figure 7-2 (b)) for the same speech stimuli containing car type noise. The metric pair chosen was MFCC-SRBIM and ESTOI, which observed a decrease in correlation magnitude as a function of increasing stretch factor. The stretch factor of 1.0 under uniform speech rate modulation had nearly perfect reconstruction of the original speech patterns as the stimuli before rate modulation. On the other hand, stretch factor 2.0 resulted in a wider spread of the points, due to a decrease in the similarity of the two metrics.

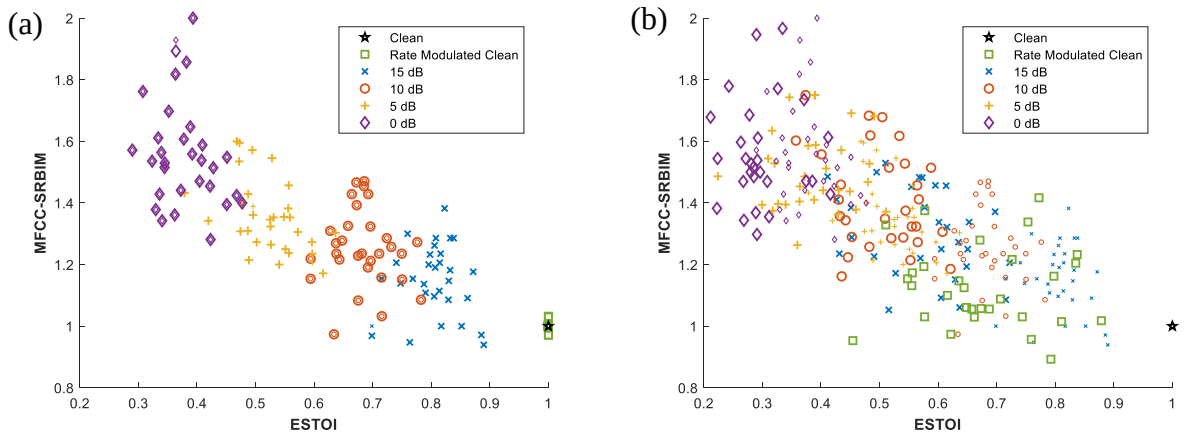


Figure 7-2: Scatter plots between MFCC-SRBIM and ESTOI using speech stimuli with car type noise for PSOLA-Uniform, stretched by (a) 1.0 times the original, and (b) 2.0 times the original. The small markers indicate the speech stimuli before uniform rate modulation, while the large markers indicate the speech after rate modulation. PSOLA-Uniform has nearly perfect reconstruction in the correlation behaviour at stretch of 1.0, with a wider variance observed at higher stretch factors

However, Figure 7-2 also shows that the metric comparisons between MFCC-SRBIM and ESTOI remained to be roughly representative of a linear relationship, as seen by the placements of the stimuli plots. The correlation magnitude also remained above 0.5, which was a fair relationship. Similar behaviour was observed with the remaining pairs of SRBIM to either ESTOI or STOI, with an increase in stretch factor resulting in greater variation from the line of best fit. However, a completely different change in the pattern of scatter was observed between MFCC-SRBIM and SNR loss with an increase in stretch factor, which observed an increase in correlation magnitudes.

Figure 7-3 shows the two scatter plots between MFCC-SRBIM to SNR loss for speech with car noise at stretch factors of 1.0 (Figure 7-3 (a)) and 2.0 (Figure 7-3 (b)). Previously, the scatter plot between SNR loss with ESTOI was noted to be better described by a nonlinear, rather than a linear relationship. This observation also applied for comparisons between SRBIM and SNR loss even after rate modulation. Especially with the scatter plot stretched at 1.0 times the original signal (Figure 7-3 (a)), the stretched signals lie along a nonlinear function that increases exponentially. The increase in stretch factor resulted in a migration of the scatter plots on this nonlinear line of best fit towards the right, resulting in a higher level of clustering in the scatter plots along the steep portion of the exponential curve (Figure 7-3 (b)).

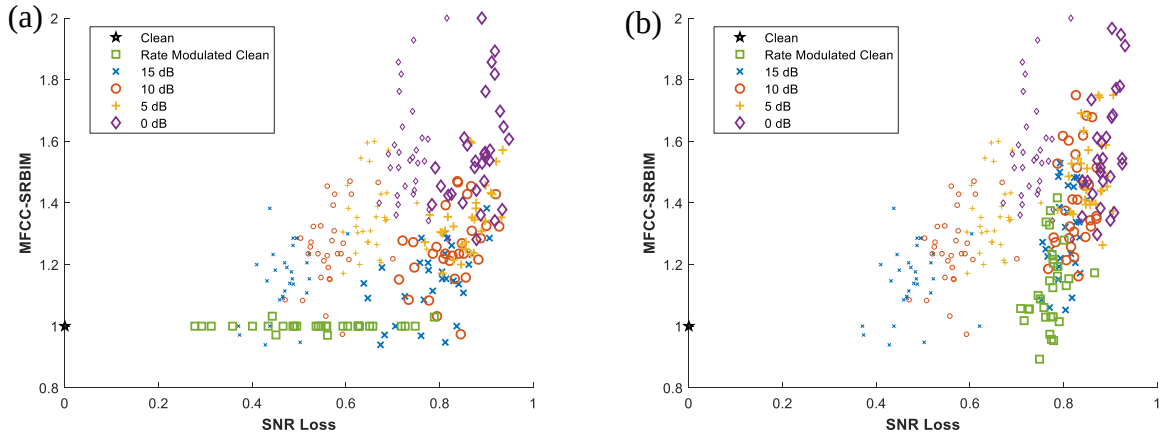


Figure 7-3: Scatter plots between MFCC-SRBIM and ESTOI using speech stimuli with car type noise for PSOLA-Uniform, stretched by (a) 1.0 times the original, and (b) 2.0 times the original

This high degree of clustering resulted in an increase in correlation magnitude in all metric pairs containing SNR loss. However, this increase in correlation was not a reflection of the validity of using SNR loss as a speech intelligibility metric for rate modulated speech. The high degree of clustering suggests that SNR loss found rate modulated speech to be equally low in speech intelligibility, which was not aligned with the other metrics, or with my subjective observations, noted in later sections. Once again, the shortcomings of using correlation coefficients were observed through these analyses of uniform speech rate modulation.

On the other hand, the existing non-uniform speech rate modulation algorithm showed more consistency in the correlation magnitudes throughout different maximum stretch factors used. Figure 7-4 shows a plot of the correlation magnitudes for the metric pairs graphed as a function of maximum stretch factor used in the existing non-uniform speech rate modulation, PSOLA-Non-Uniform. Magnitude changes for pairs involving SNR loss show a nonlinear, rather than a linear relationship with the scatter plot presented in Figure 7-3, and will be not be considered in greater detail during the remainder of the analysis. In contrast, the roughly linear relationships between SRBIM and ESTOI, and SRBIM and STOI, observed an increase in correlation from stretch factor of 1.0 to 1.2, and remained relatively constant for all other stretch factors used. This was most likely due to the nature of the stretch function used.

Chapter 2 described how the existing non-uniform speech rate modulation algorithm worked by classifying each segment of the speech under analysis as a consonant, vowel, plosive, phoneme transition or a pause, and assigned fixed stretch values to each (Demol et al., 2004). The stretch values were designed so that they were evenly incremented throughout the five groups. Additionally, an increase in maximum stretch factor was programmed to reflect as an even increase across all speech segment groups, maintaining the relative ratios between stretch factors, preserving the correlation magnitudes across most of the stretch factors. The exception, which was a large jump from maximum stretch factor of 1.0 to 1.2 in Figure 7-4, was most likely caused by the change from uniform speech rate modulation, at stretch factor of 1.0, to a non-uniform one at stretch factor 1.2. However, there was another observation that must be noted, pertaining to correlation magnitudes observed for SRBIM to ESTOI and SRBIM to STOI pairs for all three non-uniform speech rate modulation algorithms.

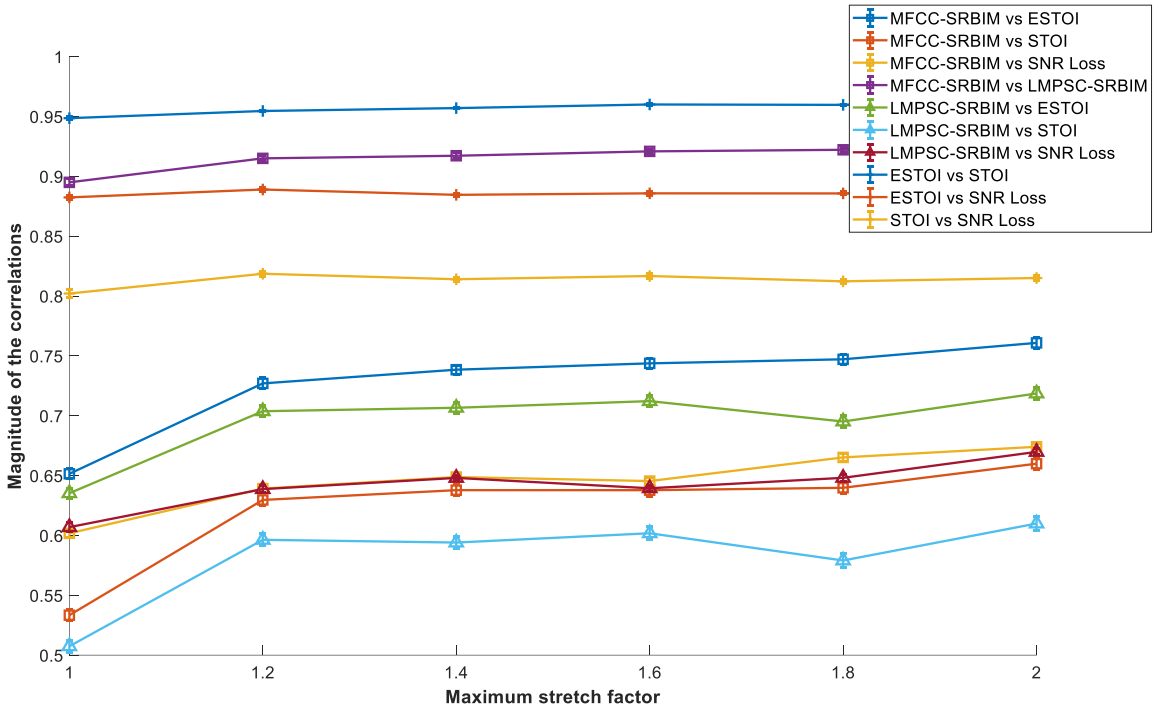


Figure 7-4: Magnitude of correlations between metric pairs for PSOLA-Non-Uniform rate modulation for car type noise. Increase in correlations between all metric pairs compared to SRBIM is observed when the maximum stretch factor is increased from 1.0 to 1.2

The magnitude of the correlations at maximum stretch factor of 1.0 between MFCC-SRBIM to ESTOI and STOI, as well as between LMPSC-SRBIM to ESTOI and STOI were in the range of 0.5 to 0.65 for the existing non-uniform speech rate modulation (Figure 7-4), as well as the two variants of SOLANU (Figure 7-5 and Figure 7-6). This was in contrast to the same metric pairs resulting in correlation magnitudes in the range of 0.8 to 0.85 with the uniform speech rate modulation (Figure 7-1). Since the minimum stretch value for all non-uniform speech rate modulations was fixed at 1.0, a maximum stretch factor of 1.0 would effectively result in uniform speech rate modulation. However, both MFCC-SRBIM and LMPSC-SRBIM, as well as ESTOI and STOI reflected a decrease in intelligibility via the non-uniform speech rate modulation algorithms that used a maximum stretch factor of 1.0. This ultimately decreased the correlation magnitudes between these metric pairs. This decrease can be explained as the difference in the introduced distortions during the analysis and synthesis stage of rate modulation.

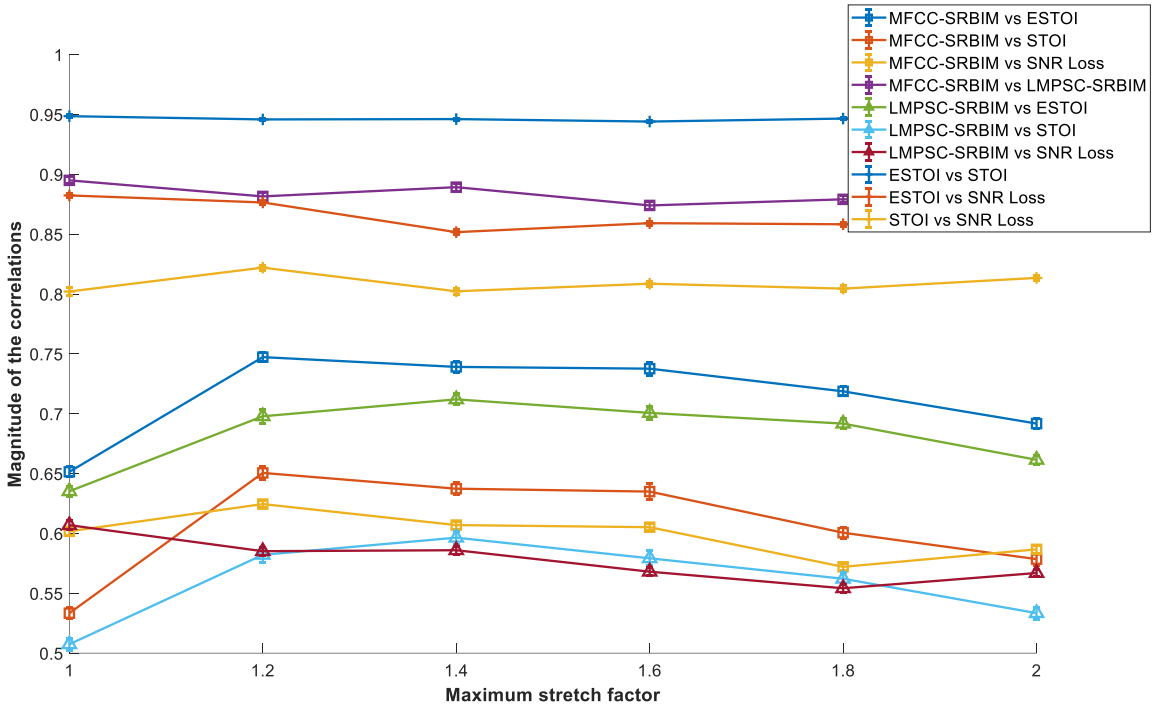


Figure 7-5: Magnitude of correlations between metric pairs for SOLANU-MFCC rate modulation for car type noise. Increase in correlations between metric pairs comparing SRBIM to ESTOI or STOI is observed when the maximum stretch factor is increased from 1.0 to 1.2, with a gradual decrease in the same metric pairs with subsequent increase in stretch factor

Both uniform and non-uniform speech rate modulation algorithms effectively decompose the speech signal into parts during analysis, before reconstructing it during the synthesis stage. As such, there is room for distortions to be introduced. Additionally, non-uniform speech rate modulation algorithms were coded with additional safety precautions such as the addition or deletion of points to allow a complete window to be sampled, as well as minor adjustments in the epoch alignment phase, which meant more opportunities for distortions to sneak in. These distortions introduced through non-uniform speech rate modulation were reflected in the objective measures of speech intelligibility. This was an interesting find, considering the fact that the non-uniformly rate modulated speech had no discernable difference from its original speech file to the human ear.

In terms of the non-uniformly rate modulated speech with a maximum stretch factor of 1.0, the author's personal observations was that there were no perceivable differences between these speech stimuli, the uniformly rate modulated version using a stretch factor of 1.0, and the

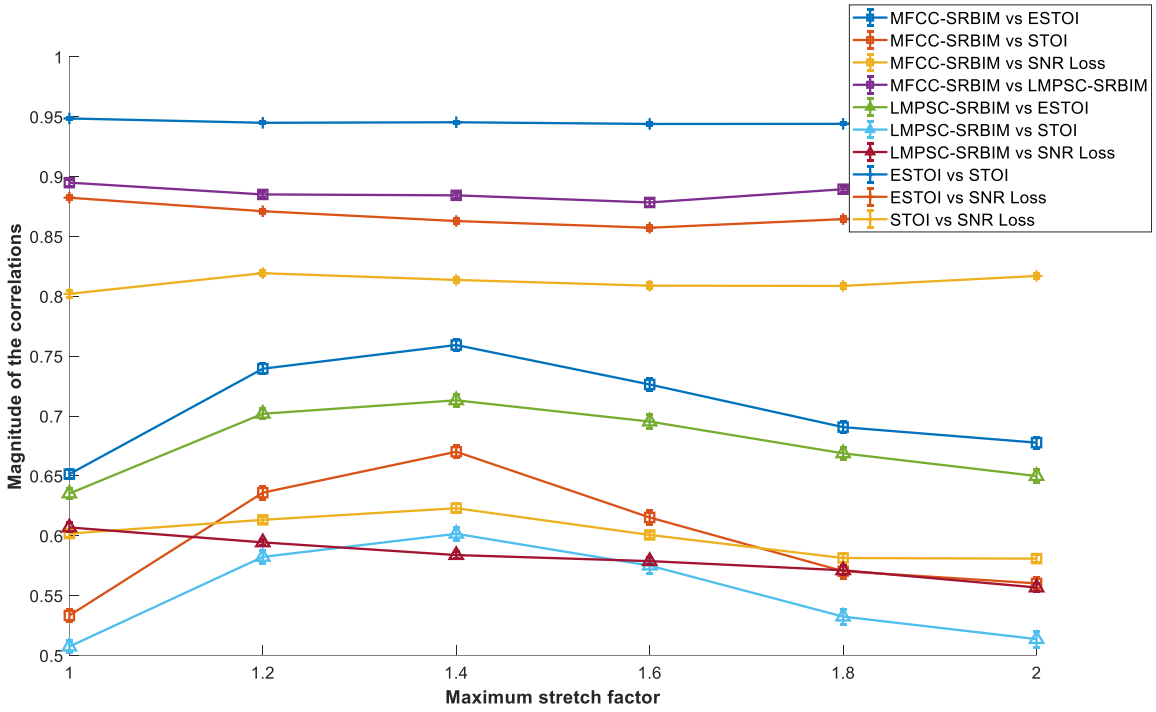


Figure 7-6: Magnitude of correlations between metric pairs for SOLANU-LMPSC rate modulation for car type noise. Increase in correlations between metric pairs comparing SRBIM to ESTOI or STOI is observed when the maximum stretch factor is increased from 1.0 to 1.4, with a gradual decrease in the same metric pairs with subsequent increase in stretch factor

original speech. Nonetheless, a separate MATLAB program confirmed the fact that while the differences were not perceivable by the human ear, the metrics observed a decrease in intelligibility. This difference may be due to the fact that I am a normal hearing and relatively young observer. Hearing impaired listeners may be able to perceive of this difference. Nonetheless, these results indicate that the objective measures are sensitive to changes in intelligibility that cannot be detected by a normal hearing observer.

An additional observation in the correlation magnitudes plotted as a function of maximum stretch value in the two variants of the proposed SOLANU algorithm (Figure 7-5 and Figure 7-6), was that the metric pairs of SRBIM to ESTOI and SRBIM to STOI seemed to peak in magnitude at stretch values of 1.2 and 1.4, relative to stretch factors on either the lower or upper end. Once again, a scatter plot gives some possible explanations to this observation.

Figure 7-7 shows the scatter plots between LMPSC-SRBIM and ESTOI for the three different maximum stretch factors of 1.0 (Figure 7-7 (a)), 1.4 (Figure 7-7 (b)) and 2.0 (Figure 7-7 (c)) used in the LMPSC variant of the SOLANU algorithm. The increase in correlation

magnitude from stretch factor of 1.0 to 1.4 was most likely the result of the migration of the plots of the rate modulated clean signals (green squares). While the clean rate modulated speech stretched by 1.0 times (Figure 7-7 (a)) was considered to have LMPSC-SRBIM on par with the original clean speech, when the speech was maximally stretched by 1.4 times (Figure 7-7 (b)), much of the clean rate modulated speech saw an increase in the SRBIM, implying a decrease in intelligibility. This change resulted in a greater magnitude of correlation as ESTOI rated these rate modulated clean signals to be much lower in intelligibility than the original clean signals for both stretch factors of 1.0 and 1.4. The correlation increased between LMPSC-SRBIM and ESTOI when the former metric mimicked the decreased intelligibility observed in ESTOI at maximum stretch factor of 1.4, realigning much of the scatters along an invisible line of best fit.

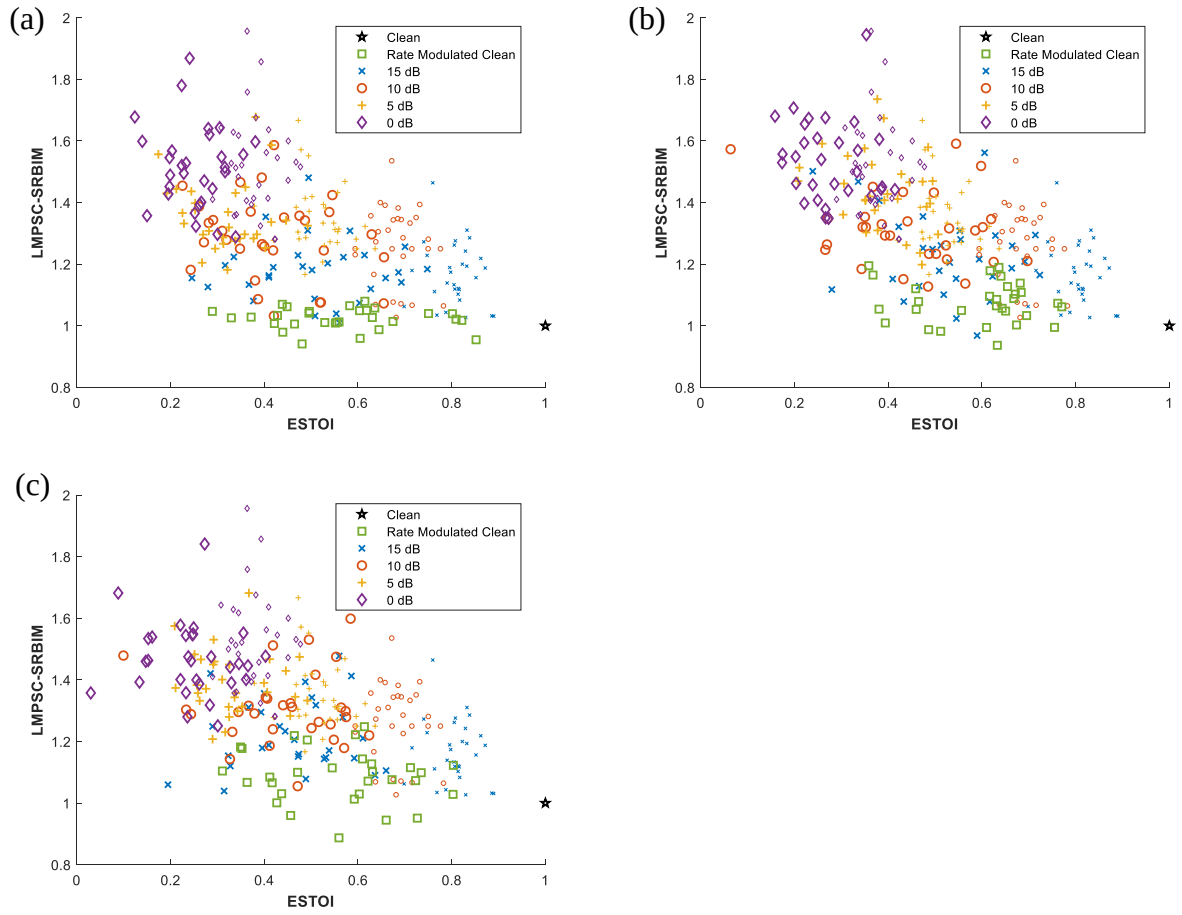


Figure 7-7: Scatter plots between LMPSC-SRBIM and ESTOI using speech stimuli with car type noise for SOLANU-LMPSC, stretched maximally by (a) 1.0 times the original, (b) 1.4 times the original, and (c) 2.0 times the original. The migration of the rate modulated clean signals (green squares) to have higher LMPSC-SRBIM at maximum stretch factor of 1.4 resulted in a higher correlation as compared to a maximum stretch factor of 1.0. Maximum stretch factor of 2.0 observes more stray scatters in both metrics, resulting in a decrease in correlation

The decrease in correlation magnitude from maximum stretch factor of 1.4 to 2.0 was a result of the larger variances in both metrics resulting in a somewhat fuzzy clump of points. At all noise levels, more stray plots away from the line of best fit are observed. However, as mentioned many times before, correlation does not directly translate to the validity of the metric performance in rating intelligibility of rate modulated speech. This is better analyzed with the inclusion of the author's personal observations on the intelligibility and quality of the rate modulated speech.

7.2 Comparison of Rate Modulated Speech with Subjective Observations

Several subjective observations were noted when listening to a collection of speech files, both clean and with noise present, in its original form and rate modulated via one of the four available methods. Whether any of the five metrics were able to detect these subjective observations was investigated in this section.

The first obvious observation was that the non-uniform speech rate modulations, regardless of the algorithm implemented, was greatly preferred over the uniform speech rate modulation, especially when the speech contained background noise. When listening to the rate modulated speech, the uniform method of rate modulation seemed to produce speech that was lethargic and unnatural. In the presence of background noise, especially babble and restaurant type noise which was highly fluctuating and contained human voices, the simultaneously rate modulated background noise was distracting, and could be described as "warbly".

Additionally, parts of words, such as the beginnings of "actress" and "smooth" seemed to disappear for uniformly rate modulated speech in the presence of noise. The resulting speech sounded like "-tress" and "-mooth". The author was able to detect that something was stated before the "-tress", but the details were lost. This phenomenon seemed to become slightly worse with an increase in stretch factor. The same word components were also noted to be slightly faint, but present, with the existing non-uniform speech rate modulation outputs, which had a faster overall presentation rate than the uniformly modulated output. Even faster resultant speech in the SOLANU variants of non-uniform speech rate modulation observed that these speech componenets were intact and perceivable, suggesting overall speech rate had a large influence on the perception of these speech cues.

While speech was noted to be unnatural and lethargic with uniform speech rate modulation, non-uniform speech rate modulation resulted in a much more natural sounding

speech, which was also generally faster. The existing non-uniform speech rate modulation (PSOLA-Non-Uniform) was perceived to be slightly faster than the uniform rate modulation, and the two variants of the SOLANU outputs were considered to have an even faster speech rate. For most speech stimuli, while the three non-uniform speech rate modulation algorithms were noted to produce outputs that varied in their overall speech rate, the difference of preference and perceived intelligibility was slight. There were certain occassions where sounds like “act” from “actress” were still slight with PSOLA-Non-Uniform, which sounded clearer with SOLANU speech. In general, my preferences seemed to be related to the overall rate of the resultant speech, with faster speech being preferred over slower speech. Slower speech was perceived to have distracting background sounds that made certain aspects of the target speech faint sounding.

However, when considering clean speech that did not have any background noise, speech rate modulation, especially non-uniform speech rate modulation, sounded as clear and intelligibile as the original stimuli. In fact, the preserved speech characteristics in conjunction to the slower speech rate, resulted in speech that often sounded quite articulate. There were some cases where the algorithm for actually time scaling the speech introduced a warbly distortion into the target voice, but for others, the speech characteristics were highly preseved. The warbly distortion was also less noticeable in the PSOLA-Non-Uniform in comparison to the PSOLA-Uniform, and even less noticeable in the two SOLANU variants, suggesting that a faster speech rate resulted in less perceivable distortions. The unnaturalness of the existing uniform speech rate modulation outputs did persist, with words like “actress” being stretched out to sound like “actressss…”.

Of these listed observations, the fact that speech rate modulation resulted in an improvement in speech intelligibility with clean speech more so than speech containing background noise was reflected in the newly proposed SRBIM. Figure 7-8 shows the normalized improvements in ESTOI (Figure 7-8 (a)), STOI (Figure 7-8 (b)) and SNR loss (Figure 7-8 (c)), previously calculated for the mock experiment comparing the performance of the three noise reduction algorithms. The same normalized improvements indicated that all three existing metrics judged that distortions were introduced by the four types of speech rate modulation algorithms, regardless of whether background noise was present or not. On the other hand, SRBIM improvements, shown in Figure 7-9, show promising results in the improvements detected for the clean speech.

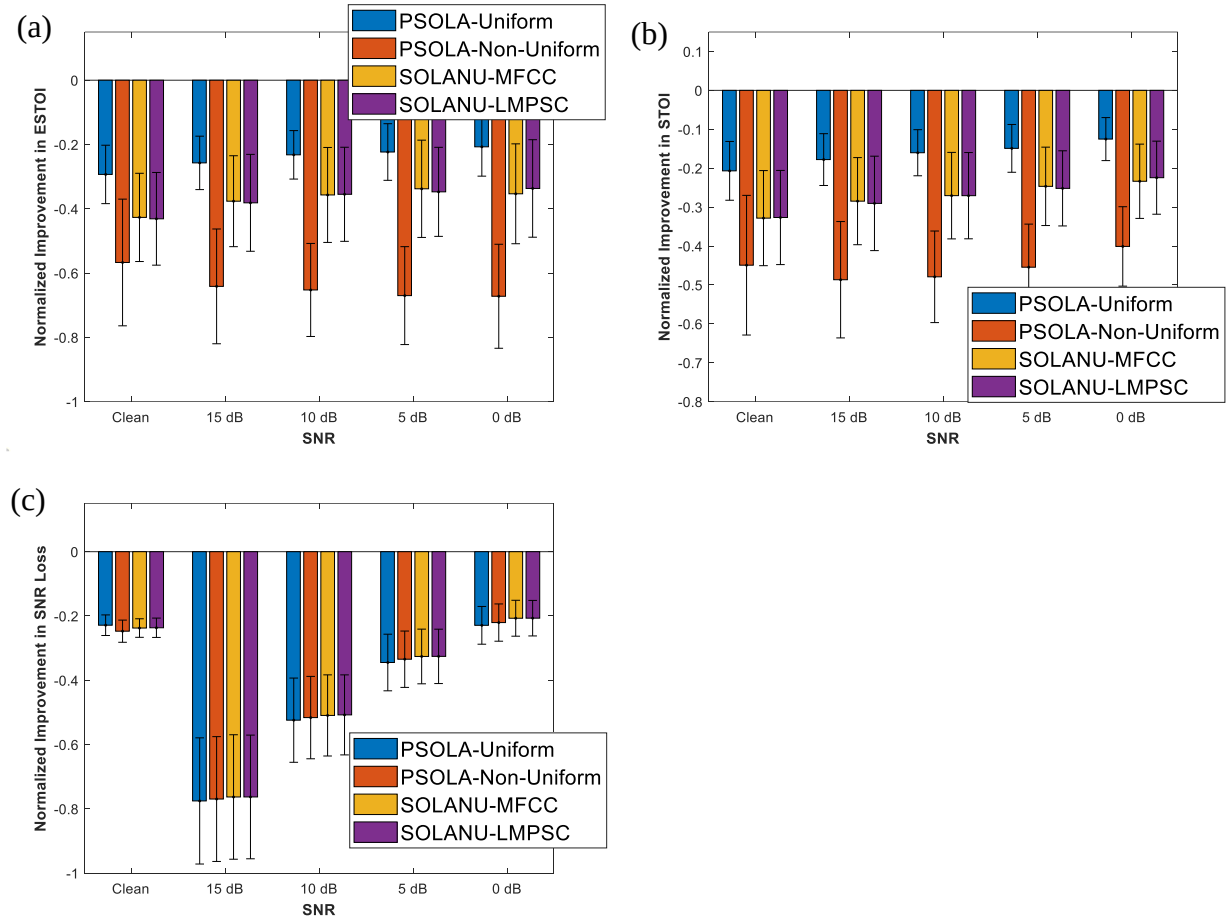


Figure 7-8: Normalized improvements in (a) ESTOI, (b) STOI, and (c) SNR loss after rate modulating speech by a factor of 1.8 using various speech rate modulation algorithms including PSOLA-Uniform (blue), PSOLA-Non-Uniform (red), SOLANU-MFCC(yellow) and SOLANU-LMPSC (purple). Error bars indicate a range of two standard deviations from the mean. All three metrics detected the changes in speech caused by rate modulation to be compromising speech intelligibility regardless of the type of rate modulation used

In other words, the author perceived the clean signals that were rate modulated, to have intelligibilities equal to or better than before rate modulation, especially with non-uniform speech rate modulation. Additionally, the noisy speech that were rate modulated were perceived to have experienced a decrease in intelligibility as compared to before rate modulation. It was evident that the clean rate modulated speech had much higher intelligibility than its noisy counterparts. Figure 7-8 shows that all three existing metrics sensed that speech intelligibility after rate modulation decreased for both clean and noisy speech. On the other hand, Figure 7-9 shows that both MFCC-SRBIM and LMPSC-SRBIM sensed a decrease in speech intelligibility after rate modulation for noisy speech, but an increase in intelligibility for clean speech. This

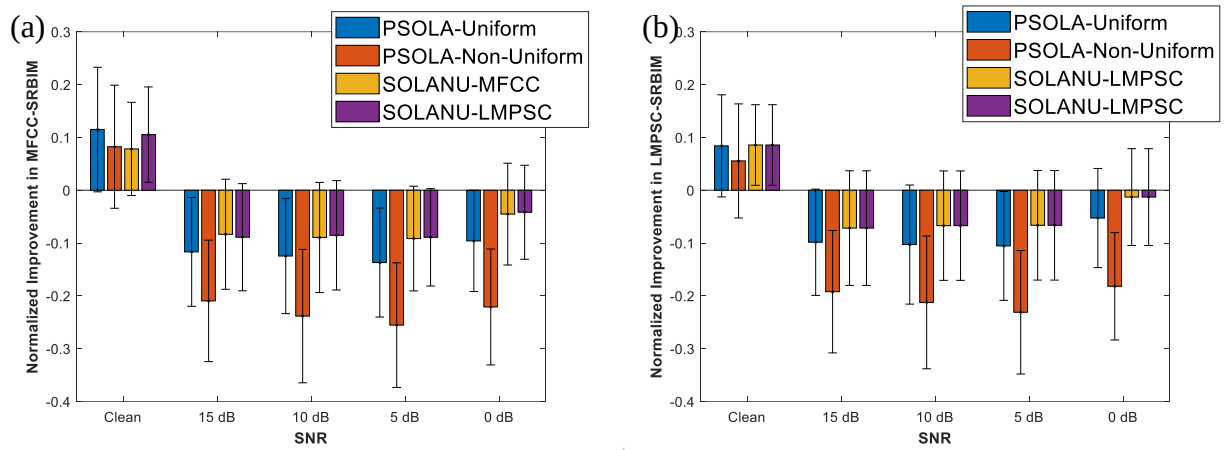


Figure 7-9: Normalized improvements in (a) MFCC-SRBIM and (b) LMPSC-SRBIM after rate modulating speech by a factor of 1.8 using various speech rate modulation algorithms including PSOLA-Uniform (blue), PSOLA-Non-Uniform (red), SOLANU-MFCC(yellow) and SOLANU-LMPSC (purple). Error bars indicate a range of two standard deviations from the mean. SRBIM showed potential in detecting the improvements in rate modulated clean speech, which were also noted through subjective observations

increase in intelligibility was reflective of the subjective observations of the clean rate modulated speech to be as clear, or perhaps even better than before rate modulation.

In summary, speech rate modulation type enhancements were subjectively rated to be beneficial only when there was no background noise present. While the three existing metrics could not successfully judge this improvement with clean speech, both MFCC-SRBIM and LMPSC-SRBIM were able to detect this slight improvement or consistency in speech intelligibility.

7.3 Summary of Simulation Results of Rate Modulated Speech

This chapter has presented the results of the simulation studies using speech that had been processed through existing uniform, existing non-uniform and two newly proposed non-uniform speech rate modulation algorithms. The correlation and scatter plot analysis in the rate modulated speech showed that there was larger variance between SRBIM and existing metrics resulting in lower correlation magnitudes. This indicated that these two metrics had started behaving differently when measuring speech intelligibility of rate modulated speech. Although there was a lower strength in the correlation, the relationship between SRBIM and ESTOI and STOI was still roughly linear.

While many subjective observations were made regarding speech intelligibility of the rate modulated speech, the most prominent phenomenon was the clarity of the clean rate modulated speech, compared to the noisy rate modulated speech. This implied that rate modulation on

simply noisy speech was detrimental to speech intelligibility. This difference in the effects of rate modulation on speech, depending on the presence or absence of noise, was only reflected in the newly proposed SRBIM. All three existing metrics in their modified form were unable to detect this difference. As such, SRBIM may have the potential of measuring differences in speech intelligibility perceived by humans with rate modulated speech, which could not be detected by existing measures.

8 Conclusions and Future Research

This final chapter summarizes the findings derived from the literature review and empirical research and connects them back to the research aim and objectives identified in Chapter 1. Recommendations for future research will also be presented.

8.1 Research Objectives: Summary of Findings and Conclusions

This thesis had a general research aim to advance the understanding of the association between speech rate and intelligibility, especially in the context of noise and noise vulnerable populations, such as the hearing impaired and elderly. This was achieved by the new SRBIM (Speech Rate Based Intelligibility Metrics) that measured speech intelligibility using speech rate information, and through the new SOLANU (SOLely Acoustic Non-Uniform) speech rate modulation algorithm that attempted to improve speech intelligibility by changing the speech rate of the original signal. The specific research objectives were as follows:

1. Identify existing objective speech intelligibility metrics, and suggest approaches by which they can be modified for use with rate modulated speech
2. Identify existing speech rate modulation based enhancement algorithms
3. Validate newly proposed metrics against existing objective speech intelligibility metrics under speech enhancement techniques which do not involve rate modulation, such as noise reduction
4. Explore the potential of newly proposed metrics and modified existing metrics in determining the speech intelligibility of the existing and newly proposed speech rate modulation based speech enhancement algorithms

The findings from the thesis pertaining to each of the research objectives will be summarized in the following sections.

8.1.1 Research Objective 1: Identify existing objective speech intelligibility metrics, and suggest approaches by which they can be modified for use with rate modulated speech

Several objective speech intelligibility metrics, and their history of improvement, were described in Chapter 2, the literature review. A comparison was made of existing metrics such as ESTOI (Extended Short-Time Objective Intelligibility), STOI (Short-Time Objective Intelligibility) and SNR (signal to noise ratio) loss, which calculated speech intelligibility by comparing the target and reference signals in the frequency domain. In their existing format,

these metrics were not able to handle speech that had be rate modulated. However new modifications to allow these existing metrics to be used in this new environment were introduced in Chapter 3.

The most important finding from conducting the literature review on existing objective speech intelligibility metrics was the fact that no previous study had attempted to objectively measure speech that had been enhanced by rate modulation. This identified gap in research was one of the issues that the thesis investigated through empirical research outlined in research objectives 3 and 4. Additionally, two variants of a new Speech Rate Based Intelligibility Metric (SRBIM) was proposed in response to this gap in literature.

Historically, objective measures of speech had been proposed out of a certain need for measures in specific situations, or as improvements upon previous measures that were documented to be inadequate in certain scenarios. Since no previous study had objectively measured rate modulated speech, this was an opportune moment to propose a speech intelligibility metric relevant to this new environment. The literature review revealed that there was a strong association between speech rate and speech intelligibility, especially for people with hearing impairments, and in situations involving noise. This was the rationale for proposing the use of objective measures of speech rate as an alternative metric for speech intelligibility. In Chapter 3, it was discussed that the new normalization step inside SRBIM would allow target and references speech to be compared easily, regardless of a mismatch in the time domain. This newly proposed SRBIM required validation obtained through empirical research.

8.1.2 Research Objective 2: Identify existing speech rate modulation based enhancement algorithms

Another identified gap in literature was the lack of well documented non-uniform speech rate modulation algorithms that could be easily replicated. People with hearing impairments and elderly listeners have difficulty comprehending speech in noise, and current assistive hearing technologies are inadequate in improving the situation. A possible solution was in people's natural tendency to slow speech down in a non-uniform manner. The literature reviewed in Chapter 2 and 3 supported the claim that slowing down speech would act as speech enhancement for hearing impaired listeners. Chapter 2 detailed several tools that are currently available to uniformly slow speech down, with emphasis placed on the popular PSOLA (Pitch Synchronous Overlap and Add) method. The difference between uniform and non-uniform speech rate

modulation was highlighted to be the use of a constant stretch factor to slow speech in uniform approaches, as compared to a time-varying stretch function for non-uniform methods. When speaking in noise, people slow down in a non-uniform manner. Additionally, listeners perceive uniformly rate modulated speech to be unnatural sounding. In other words, non-uniform speech rate modulation algorithms were expected to be better than uniform approaches as they generated speech that was more similar to the natural mannerisms of human speech, making the output easier to listen to.

However, many of the non-uniform speech rate modulation algorithms were poorly documented, making it difficult to experimentally replicate. This gap in literature led to the proposal of the new SOLely Acoustic Non-Uniform (SOLANU) speech rate modulation algorithm.

The new SOLANU algorithm, detailed in Chapter 3, was developed out of a need for a non-uniform speech rate modulation algorithm that produced natural sounding output that could be replicated in future studies. The approach was based on a well studied parameter in speech studies, Mel Frequency Cepstral Coefficients (MFCC) and its intermediate, Log Mel Power Spectral Coefficients (LMPSC). By deriving the Spectral Transition Measure (STM) of these coefficients, the boundary points in speech, where the spectral features were changing quickly, could be identified. The location of the speech boundaries was used to derive a new stretch function, where the transients were preserved by assigning those points a stretch value of 1.0. Speech segments between these identified boundaries were assigned a stretch value that increased gradually, resulting in a new modified triangular wave like stretch function. The new algorithm was included in the simulation outlined in research objective 4.

8.1.3 Research Objective 3: Validate newly proposed metrics against existing objective speech intelligibility metrics under speech enhancement techniques which do not involve rate modulation, such as noise reduction

There was a need to validate the newly proposed speech intelligibility metrics. While SRBIM was envisioned to be used for measuring intelligibility of rate modulated speech, the metrics were first compared against the existing metrics of ESTOI, STOI and SNR loss in measuring intelligibility of speech that had been enhanced through popular methods. Three types of noise reduction algorithms were chosen as common methods of speech enhancement, and the

new and existing metrics were analyzed for their correlation coefficients under situations where they measured speech intelligibility of noisy speech, as well as noise reduced speech.

Results of the simulations revealed that the new metrics had a roughly linear relationship with ESTOI and STOI, presenting itself as a valid measure of intelligibility for both noisy and noise reduced speech. SRBIM showed great potential, especially when the data was analyzed in a larger sample, consisting of only one type of noise. The metrics were also used in a mock experiment where the performance of the three noise reduction algorithms was compared through the improvements in intelligibility as measured through each metric. SRBIM showed high agreement with existing metrics in the superiority of the Wiener filter based design, but had trouble ranking the other two. However, even existing metrics were greatly conflicted in the ranking between the two similar noise reduction algorithms, supporting the notion that SRBIM may be able to confirm the ranking of largely differing options. This may be a valid method of giving directionality to the development of speech enhancement algorithms in the initial stages of implementation.

8.1.4 Research Objective 4: Explore the potential of newly proposed metrics and modified existing metrics in determining the speech intelligibility of the existing and newly proposed speech rate modulation based speech enhancement algorithms

Finally, the thesis set out to validate the newly proposed SRBIM, and the modified versions of the existing metrics, for measuring speech that had been rate modulated, the situation which was identified as a gap in literature through Chapter 2. In addition to the correlation based analysis implemented with the noise reduction algorithms, my subjective observations were included in the analysis.

The most exciting find in this scenario was regarding the metric behaviours in quantifying improvements or distortions in clean and noisy speech. One of the most noticeable differences in the rate modulated speech signals was the fact that clean rate modulated speech was much easier to comprehend than noisy rate modulated speech. The newly proposed SRBIM were the only metrics that judged rate modulation of clean speech to have improved in intelligibility. All three modified existing metrics detected that speech intelligibility decreased with the application of rate modulation regardless of whether it contained noise or not. This observation highlighted the fact that the new SRBIM may indeed have an advantage over existing metrics in measuring the speech intelligibility of rate modulated speech, the scenario in

which it was initially designed for, as it was the only metric that seemed to align with preliminary human observations.

This section had highlighted some of the findings of this thesis through its four research objectives. A summary of the thesis chapters is provided below.

8.2 Summary

This thesis had set out to advance the understanding between speech rate and speech intelligibility by proposing new metrics based on speech rate to measure intelligibility, suggesting new methods of non-uniform speech rate modulation to improve intelligibility, and a series of simulations to test them out with noisy speech, noise reduced speech and rate modulated speech. Performances of the newly proposed metrics were compared to three existing metrics through correlation coefficients, scatter plots as well as a mock experiment for noise reduced speech and some preliminary subjective observations for rate modulated speech.

Chapter 1 provided a general background of normal hearing listener's natural ability to extract relevant speech from noise, which was a skill that was lacking in elderly or hearing impaired listeners. Current assistive hearing technologies were unhelpful in improving the comprehension of speech in the presence of noise. As a possible solution, the chapter concluded with people's natural tendency to slow down speech.

Chapter 2 was a literature review of existing metrics and existing methods of rate modulation, which identified two key gaps in the literature. It started by providing support from literature on the benefits of rate modulating speech for better comprehension. Existing methods of rate modulation were described, noting that the literature on well documented non-uniform speech rate modulation algorithms that contained sufficient details for experimental replication, were scarce, even though this was the natural pattern of rate modulation in humans. This was the first gap in literature, identifying a need for a well documented non-uniform speech rate modulation algorithm.

Chapter 2 identified another gap in research pertaining to objective speech intelligibility measures. Namely, no previous study had attempted to objectively measure the intelligibility of speech that had undergone speech rate modulation. There were many objective intelligibility metrics that had been proposed for use in specific situations, but no previous study had used these measures for rate modulated speech. In fact, many of the existing metrics requested perfect time alignment between the reference and target signals. As such, there was a need for new

metrics and modified versions of existing metrics that could be used to handle target speech signals that had undergone rate modulation, in calculating its intelligibility.

Chapter 3 described the modifications to existing intelligibility metrics to make them capable of handling rate modulated speech, and newly proposed algorithms for measuring and improving speech intelligibility. Existing speech metrics that compared reference and target speech in the frequency domain could theoretically be modified to compared two speech signals that differed in its rate, by changing the window length and step size used during analysis. Additionally, the strong association between speech intelligibility and speech rate, introduced in Chapter 2, hinted at the possibility of using objective measures of speech rate to measure speech intelligibility. This was the basis for the proposal of Speech Rate Based Intelligibility Metrics (SRBIM), which used speech boundaries detected through STM-MFCC and STM-LMPSC to calculate absolute speech rate of both target and reference signals. These absolute rates were normalized by taking the ratio between the target and reference speech rates, to yield the two variants of SRBIM, MFCC-SRBIM and LMPSC-SRBIM.

Chapter 3 proposed a new method of improving speech intelligibility through SOLely Acoustic Non-Uniform (SOLANU) speech rate modulation. Two variants of SOLANU, SOLANU-MFCC and SOLANU-LMPSC were also introduced, based on speech boundaries detected through STM-MFCC and STM-LMPSC, respectively. The heart of this new algorithm was in the derivation of the stretch function, where a modified triangular wave was suggested and tested. The theory behind SOLANU's time-varying stretch function was to preserve the timing information at the speech boundary points, which were identified as points of sudden change in the spectral or cepstral coefficients, by assigning these segments a stretch factor of 1.0. Speech segments between the identified speech boundaries were rate modulated to be slower through the assignment of an incrementally changing stretch factor. The resultant stretch function of SOLANU looked like a series of triangular waves.

Chapter 4 described the details of the research methodology used in the series of simulation experiments that compared the newly proposed metrics against the existing metrics in scenarios involving simply noisy speech, and speech enhanced through noise reduction and rate modulation. This included the methodology on data collection and data analysis. Simulations using the NOIZEUS database (Hu & Loizou, 2007) were chosen as this would allow for incremental testing using the same target speech stimuli, differing slightly in levels of noise, or

the type of speech enhancement applied. This type of incremental testing would not have been possible with human observers. The chapter also provided the methodology of generating noise reduced versions and rate modulated versions of the speech files from the original noisy and clean speech signals. Three types of noise reduction algorithms, the geometric approach to spectral subtraction, the probabilistic geometric approach to spectral subtraction and the two-stage mel-warped Wiener filter based noise reduction algorithms, were used. For rate modulation, an existing uniform, an existing non-uniform and two variants of the newly proposed SOLANU speech rate modulation algorithms, were used. All simulations analyzed the metric performances through correlation coefficients and an analysis of the scatter plots between SRBIM and existing metrics. These were compared to the correlation and scatter observed within existing metric pairs to evaluate SRBIM performance. The agreement rate between the new and existing metrics during the mock experiment, where metrics independently ranked the performances of the three noise reduction algorithms, added another dimension of analysis. Lastly, for the simulation involving rate modulated speech, my personal observations were included to investigate whether any of the metrics reflected differences perceivable to the human ear.

The findings of the simulations were spread over Chapters 5, 6 and 7. Chapter 5 presented the results pertaining to the optimization of the two unknown parameters required in SRBIM calculations, as well as the findings of the metric comparisons for simply noisy speech. Chapter 6 described the findings pertaining to the two new and three existing metrics for speech that was noise reduced through the three noise reduction algorithms. Analysis in Chapter 6 also included a mock experiment where the metrics were used to rank the noise reduction algorithm performances. Chapter 7 showed the results associated with metric comparisons in evaluating the intelligibility of rate modulated speech.

Chapter 5 started with the results involved in arriving at the optimized parameters required for producing SRBIM outputs that behaved as expected. It was found that SRBIM was best calculated with a threshold factor of 1.0 in order to observe an increase in SRBIM, or a decrease in intelligibility, with an increase in noise levels. On the other hand, window length, over which the adaptive median was calculated, had a much smaller effect on SRBIM outputs. Nonetheless, the window length that produced the largest magnitude in the correlations was used in the remaining simulations, which was 90 ms for MFCC-SRBIM and 110 ms for LMPSC-SRBIM.

The results pertaining to noisy speech in Chapter 5 showed that SRBIM was linearly related to ESTOI and STOI, given a large enough sample, and especially when tested with one type of noise. The correlation coefficients showed that this relationship was relatively strong, and the distribution of the scatter plots supported the fact that the relationship was linear. On the other hand, SNR loss, an existing metric, was found to be nonlinearly related to all other metrics. These results suggest that SRBIM had potential in appropriately quantifying the intelligibility of noisy speech when analyzed as a collection of samples.

Chapter 6 presented the metric behaviours for noise reduced speech. The results from the correlation and scatter plots were similar to those observed in Chapter 5. Namely, given a large enough set of samples, MFCC-SRBIM and LMPSC-SRBIM showed a relatively strong linear relationship with ESTOI and STOI. The mock experiment also showed the SRBIM had high agreement with existing metrics when a large sample size was used, especially in higher levels of noise. At these levels, many of the exiting metrics and SRBIM agreed on the superiority of the Wiener filter based design. When less noise was present, even the existing metrics had difficulty agreeing with each other on the ranking of the noise reduction performances. These results suggested that while SRBIM was designed specifically for the situation involving rate modulated speech, it was also capable of measuring speech intelligibility of noisy, as well as noise reduced speech.

Chapter 7 presented the results on the metric behaviours as they measured intelligibility of speech that had been rate modulated through two existing and two newly proposed methods. While correlation and scatter plots suggested that SRBIM observed a larger variance in its linear relationship with ESTOI and STOI, strength of the correlation was maintained to values above 0.5, which would be considered a fair relationship. Additionally, inclusion of the preliminary subjective observations revealed that SRBIM may be the only metric capable of detecting the difference in intelligibilities between clean and noisy speech that had undergone rate modulation. Specifically, I had observed that while noisy rate modulated speech decreased in intelligibility, regardless of the rate modulation implemented. At the same time, I observed that clean rate modulated speech had equal or better intelligibility than before rate modulation. SRBIM were the only metrics that showed potential in detecting the benefits of rate modulating clean speech, as compared to noisy speech. Existing metrics labeled all changes in the rate modulated speech as distortions rather than improvements.

8.3 Recommendations for Future Research

This section gives recommendations on extending this research by using elements from speech rate analysis to improve and measure speech intelligibility, as well as recommendations for the broader scope of improving speech intelligibility for people who are particularly vulnerable to the effects of noise.

8.3.1 Inclusion of Human Observers

The research methodology used in this thesis has some limitations in its design. Much of these limitations were engrained in the chosen design, and a decision had to be made regarding whether incremental testing or real-world applicability was desired. In this section, two prominent issues regarding the use of correlation coefficients and the lack of subjective speech intelligibility scores will be covered. The use of human observers and using correlation analysis between SRBIM and human ratings may provide the opportunity to resolve some of these issues.

SRBIM performance was evaluated through Pearson correlation coefficients. These correlations range from a value of -1 and 1 and give an indication of the strength and directionality of the relationship between two metrics. Direction is judged easily by observing whether the correlation is a positive or a negative value. However, as mentioned before, the strength of the similarity between the two metrics, based on the magnitude of the correlation, can be difficult to judge. The analysis of the simulations in this thesis was based on the general guideline that correlations between 0 and 0.25 to have little to no relationship, 0.25 to 0.5 to be considered a fair relationship, 0.50 to 0.75 to be in the range of moderate to good relationship, and anything above 0.75 to constitute as good to excellent (Portney & Watkins, 2015).

However, throughout Chapters 5, 6 and 7, the correlation coefficients between SNR loss and ESTOI, and between SNR loss and STOI have been observed to be in the moderate to good range. Upon analyzing the scatter plots, it was concluded that SNR loss exhibited a nonlinear relationship with the other metrics. Since correlation coefficients evaluate a linear dependence between the two metrics, it cannot quantify nonlinear relations well (Portney & Watkins, 2015). As such, the high magnitude of correlation observed with SNR loss was misleading to its validity and strength of similarity to ESTOI and STOI. Conversely, while SRBIM scored lower correlation magnitudes, a scatter plot revealed the relationship with ESTOI and STOI to be roughly linear.

Additionally, SRBIM were evaluated by correlations with other existing speech intelligibility metrics. The assumption was that ESTOI, STOI and SNR loss were adequate estimates of actual intelligibility, and the strength of the correlation would give an indication of SRBIM ability to measure intelligibility. However, even if SRBIM was evaluated to have a moderate relationship with existing measures, whether this could be generalized to speech intelligibility in the desired population was unknown. In order to evaluate SRBIM ability to estimate intelligibility corresponding to a certain group of listeners, a sample of speech intelligibility ratings from human observers, preferably from those with a hearing status or age range appropriate to the target population, would be required.

Human testing was not included extensively in this thesis, except for brief observational notes made by the author for rate modulated speech. The main reason was that including human testing would require a more extensive speech database, and the number of noise conditions and tested speech enhancement programs would have to be limited so that the presented number of stimuli was not excessive. For example, while simulation experiments using objective measures allows comparison of speech enhancement via rate modulation at six stretch factors ranging from 1.0 to 2.0, this may have to be reduced by half to three factors for subjective measures.

Additionally, a whole new speech database would have to be generated, containing a greater number of speech stimuli. This can be recorded under controlled conditions, much like how NOIZEUS was prepared (Hu & Loizou, 2007), or recorded from live conversations where background noise is already present. If live conversations are used, one must take care in evaluating speech intelligibility after enhancement. As the same speech stimuli cannot be presented twice to the same listener, different groups of stimuli would need to be enhanced through a select number of methods. Additionally, the choice of stimuli and enhancement combination may have an effect, which can be reduced by mixing the stimuli groups with different speech enhancement protocols with each new human observer. If there is an association between noise type and enhancement model, an analysis of the noise characteristics present in the speech is desired for further discussion. However, as it would be difficult to record only the noise characteristics in real-life recordings, where there are multiple sources of noise, and speech processing of noisy speech changes the noise characteristics in the output speech altogether, it can be difficult to pinpoint the cause of the observed phenomena using real-life audio.

In future research, it would be possible to generate this appropriate speech stimuli and implement human testing. Especially as the preliminary results from this thesis indicate that there is a large difference between the intelligibilities of rate modulated speech that contains no background noise, and those that do, the focus may be placed on these two groups of stimuli. Additionally, it would be interesting to investigate the effects of noise reduction before rate modulation to see whether human listeners perceive this to be beneficial or compromising to speech intelligibility. Lastly, since SRBIM was designed based on the theory that speech rate and intelligibility are closely linked in people with hearing impairments, the recruitment of human observers may have to consider the hearing status of the subjects.

The inclusion of human observers is a recommendation of future research that directly extends upon this thesis. Another recommendation directly linked to the thesis is pertaining to the shape of the stretch function used in SOLANU.

8.3.2 Varying the Shape of the Stretch Function

SOLANU, the newly proposed non-uniform speech rate modulation algorithm, used a modified triangular wave as its stretch function in this thesis. It would be interesting to investigate the effects of stretch function shape on the generated rate modulated speech.

For example, the existing non-uniform speech rate modulation algorithm, effectively uses a stretch function shaped much like a modified version of a square wave (Demol et al., 2004). This algorithm classified speech segments into one of five classes, and assigned a fixed stretch factor depending on this classification. The resulting stretch function had sections that were rate modulated by a constant stretch factor, which would suddenly jump between sections. The triangular stretch function was an attempt to alleviate these sudden jumps in rate modulation.

Another non-uniform speech rate modulation algorithm implemented a sawtooth type stretch function. In this model, spectral transition points, or speech boundaries, were also used to help shape the stretch function. Specifically, once a speech boundary was detected, the stretch function as designed to increase suddenly, before decreasing gradually, until the next transient was detected (Grofit & Lavner, 2008).

Since previous non-uniform speech rate modulation algorithms are using a variety of stretch function shapes, it may be interesting to test different synthetically generated stretch functions in future studies. Inclusion of subjective perception of the various speech, generated by the different stretch functions may also yield some interesting finds regarding the recommended

speech rates at speech boundaries, which are assumed to hold the crucial speech cues required for comprehension. Namely, different stretch function shapes may generate rate modulated speech with different intelligibilities. Analysis of the differences in the rate modulation patterns may reveal a consistent pattern in rate modulation regarding these speech cues in order to generate highly intelligible output.

Alternatively, future studies may start with the analysis in human speakers of the differences in speech rate patterns between speech spoken in quiet and speech spoken in noise. As humans have a natural tendency to slow down speech in noise, understanding the pattern of rate modulation may lead to the development of a stretch function derivation approach that produces rate modulated speech that mimic these natural mannerisms used in human speech.

The investigation of the pattern of the stretch function and its effects on the intelligibility is a direct extension of the thesis topic investigated. The following are recommendations for research with the more general aim of improving speech intelligibility of noisy speech for people with hearing impairments.

8.3.3 Auditory Training to Improve Hearing Ability

While this thesis took the approach of improving speech intelligibility by modifying the speech, there are methods to improve comprehension by improving the listener's capabilities. Hearing loss can cause changes in the brain that makes speech comprehension even more challenging (Tremblay et al., 2003). Conversely, if depriving the brain of speech signals induces central changes in the brain, the provision of new stimuli via hearing aids or cochlear implants may result in changes as well. In explaining the biological processes behind speech comprehension, it was mentioned that the auditory processes in the brain of an older or hearing impaired individual may be slower, or have different temporal-spectral capabilities compared to normal hearing listeners. The purpose of auditory training would be to condition the newly used brain pathways to recognize the new neural responses in a meaningful manner.

In line with these inferences, although the capacity for change and retention can be quite variable, auditory training does indeed seem to improve an individual's perception to speech cues, such as voice onset time (Tremblay, 2015). For example, in a gap detection threshold detection task, it was found that specialized auditory training resulted in a significant decrease in the detectable gaps presented in the stimuli for both older and younger adults (Kishon-Rabin, Avivi-Reich, & Roth, 2013). This is a simple task that aims to measure the temporal resolution

that can be reliably discriminated by listeners. Subjects that have higher resolution are less prone to missing crucial acoustic cues necessary in comprehending speech. In another study, it was found that auditory cognitive training can improve auditory processing speed and speech processing as tested with time compressed speech and speech perception in noise, in subjects who have experienced heart failure, a population especially susceptible to cognitive impairment (Athilingam, Edwards, Valdes, Ji, & Guglin, 2015). Specifically, the group undergoing auditory training showed an improvement in understanding time compressed speech at 65% compression ratio, and speech perception in noise at 4 dB signal to noise ratio, while the control group showed a decline in both test performances over the course of the study (Athilingam et al., 2015).

Neurologically, training changed the auditory induced responses in the brain, as well as the behavioural outcome (Anderson, White-Schwoch, Choi, & Kraus, 2013; Anderson, White-Schwoch, Parbery-Clark, & Kraus, 2013; Song, Skoe, Banai, & Kraus, 2012; Tremblay & Kraus, 2002). Neural responses to formant transition regions, segments of speech that undergoing the most rapid changes, were initiated earlier than before training (Anderson, White-Schwoch, Parbery-Clark, et al., 2013). Additionally, while the presence of background noise tends to push back the neural timing of these responses, the subjects that underwent auditory cognitive training experienced less of a timing shift compared to the control group (Anderson, White-Schwoch, Parbery-Clark, et al., 2013). This decreased change in the response latency to noisy speech supports the notion that training improves the auditory processing capability of individuals on a neural level.

In another study, older adults with hearing loss had a lopsided neural representation of speech with the slowly varying speech envelope being overly enhanced at the cost of the rapidly changing temporal fine structures, as compared to normal hearing individuals (Anderson, White-Schwoch, Choi, et al., 2013). After appropriate auditory training, not only did speech perception improve, but the neural representation more closely matched those of a normal hearing subject (Anderson, White-Schwoch, Choi, et al., 2013). Similarly, the auditory brainstem responses to a syllable stimuli changed from before and after auditory training in quiet and in noise (Song et al., 2012). This study found that training was especially effective in attuning the brain to pitch-related cues, such as the fundamental frequency, which are known to be time variable during formant transition regions (Song et al., 2012). These formant transition regions are believed to be

the regions that hold crucial speech cues in constructing meaning, and not being able to properly understand these segments, perhaps due to masking noise, can result in poorer speech perception.

The benefits of auditory training have also been observed in hearing aid users. Participants newly fitted with hearing aids for treatment of sensorineural hearing loss were presented with short speech stimuli consisting of a consonant and a vowel amidst noise over eight weeks (Stecker et al., 2006). After training, the percentage of nonsense syllable correctly identified improved, which were separate from the immediate effects gained from amplification (Stecker et al., 2006). Moreover, the benefits of training was also observed in experienced hearing aid users, and was retained for at least eight weeks after training in the newly fitted group (Stecker et al., 2006).

A more practical example can be illustrated with LACE^{TM}, Listening and Communication Enhancement, a home-based interactive adaptive computer program aimed to improve comprehension of degraded speech, enhance cognitive skills and advise the user on effective communication strategies (Sweetow & Sabes, 2006). The exercises included speech recognition in the presence of a single or multiple competing speakers; comprehending rapid, time compressed speech; auditory working memory task involved in repeating the word prior to an identified target word; and a missing word exercise where the user must identify the word that is completely masked by noise in a spoken sentence (Sweetow & Sabes, 2006). While it was found that the trained group improved in all on-task measures over the course of the training period, what was more interesting was the fact that their off-task measures, such as objective measures in a speech recognition in noise, improved (Sweetow & Sabes, 2006). Moreover, these improvements persisted at four weeks after the training ended (Sweetow & Sabes, 2006).These results seem to indicate that targeted auditory training does improve the overall hearing ability and experience of people with hearing impairments, and that these benefits are translatable to situations encountered in daily life.

Several studies support the benefits of auditory training on speech comprehension. Especially with the great reluctance of hearing aid owners to use their devices, research into effective methods of training may be an area of future research with great interest from the affected individuals.

References

Aharonson, V., Aharonson, E., Raichlin-Levi, K., Sotzianu, A., Amir, O., & Ovadia-Blechman, Z. (2017). A real-time phoneme counting algorithm and application for speech rate monitoring. *Journal of Fluency Disorders*, *51*, 60–68. https://doi.org/10.1016/j.jfludis.2017.01.001

Anderson, S., White-Schwoch, T., Choi, H. J., & Kraus, N. (2013). Training changes processing of speech cues in older adults with hearing loss. *Frontiers in Systems Neuroscience*, *7*(November), 1–9. https://doi.org/10.3389/fnsys.2013.00097

Anderson, S., White-Schwoch, T., Parbery-Clark, A., & Kraus, N. (2013). Reversal of age-related neural timing delays with training. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(11), 4357–4362. https://doi.org/10.1073/pnas.1213555110

Athilingam, P., Edwards, J. D., Valdes, E. G., Ji, M., & Guglin, M. (2015). Computerized auditory cognitive training to improve cognition and functional outcomes in patients with heart failure: Results of a pilot study. *Heart and Lung: Journal of Acute and Critical Care*, *44*(2), 120–128. https://doi.org/10.1016/j.hrtlng.2014.12.004

Bacon, S. P., Opie, J. M., & Montoya, D. Y. (1998). The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds. *Journal of Speech, Language, and Hearing Research*, *41*(3), 549–563. https://doi.org/10.1044/jslhr.4103.549

Beck, D. L., Larsen, D. R., & Bush, E. J. (2018). Speech in noise: hearing loss, neurocognitive disorders, aging, traumatic brain injury and more. *Journal of Otolaryngology-ENT Research*, *10*(4), 199–205. https://doi.org/10.15406/joentr.2018.10.00345

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. B. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 1035–1047. https://doi.org/10.1109/TSA.2005.851998

Boldt, J. B., & Ellis, D. P. W. (2009). A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation. *European Signal Processing Conference (EUSIPCO)*, 1849–1853. Retrieved from https://ieeexplore-ieee-org.ezproxy.lib.ucalgary.ca/stamp/stamp.jsp?tp=&arnumber=7077636

Brumm, H., & Zollinger, S. A. (2011). The evolution of the Lombard effect: 100 years of

psychoacoustic research. *Behaviour*, *148*(11–13), 1173–1198. https://doi.org/10.1163/000579511X605759

Cherry, E. C. (1953). Some experiments on the recognition in speech with one and two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. Retrieved from http://www.ee.columbia.edu/~dpwe/papers/Cherry53-cpe.pdf

Cho, S., Yu, J., Chun, H., Seo, H., & Han, W. (2014). Speech Perception in Older Listeners with Normal Hearing: Conditions of Time Alteration, Selective Word Stress, and Length of Sentences. *Korean Journal of Audiology*, *18*(1), 28–33. https://doi.org/10.7874/kja.2014.18.1.28

Christiansen, C., Pedersen, M. S., & Dau, T. (2010). Prediction of speech intelligibility based on an auditory preprocessing model. *Speech Communication*, *52*(7–8), 678–692. https://doi.org/10.1016/j.specom.2010.03.004

Cooke, M., & Aubanel, V. (2017). Effects of linear and nonlinear speech rate changes on speech intelligibility in stationary and fluctuating maskers. *The Journal of the Acoustical Society of America*, *141*(6), 4126–4135. https://doi.org/10.1121/1.4983826

Cooke, M., Mayo, C., & Villegas, J. (2014). The contribution of durational and spectral changes to the Lombard speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *135*(2), 874–883. https://doi.org/10.1121/1.4861342

David, D., & Werner, P. (2016). Stigma Regarding Hearing Loss and Hearing Aids : A Scoping Review. *Stigma and Health*, *1*(2), 59–71. https://doi.org/10.1037/sah0000022

Davis, A., McMahon, C. M., Pichora-Fuller, K. M., Russ, S., Lin, F., Olusanya, B. O., … Tremblay, K. L. (2016). Aging and Hearing Health: The Life-course Approach. *Gerontologist*, *56*(S2), S256–S267. https://doi.org/10.1093/geront/gnw033

Davis, S. B., & Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366. https://doi.org/10.1109/TASSP.1980.1163420

De Leon, F., & Martinez, K. (2012). Enhancing timbre model using MFCC and its time derivatives for music similarity estimation. *European Signal Processing Conference (EUSIPCO)*, 2005–2009. Retrieved from https://ieeexplore-ieee-org.ezproxy.lib.ucalgary.ca/stamp/stamp.jsp?tp=&arnumber=6333837

Demol, M., Struyve, K., Verhelst, W., Paulussen, H., Desmet, P., & Verhoeve, P. (2004). Efficient Non-Uniform Time-Scaling of Speech with WSOLA for CALL applications. *InSTIL/ICALL Symposium*, 1–4. Retrieved from https://www.isca-speech.org/archive_open/archive_papers/icall2004/iic4_007.pdf

Dhivya, R. J., Senthamizh, S. R., & Suresh, G. (2016). Transform Based Speech Enhancement using Firefly Algorithm. *Advances in Natural and Applied Sciences*, *10*(9), 361–366.

Du, A., Lin, C., & Wang, J. (2014). Effect of speech rate for sentences on speech intelligibility. *IEEE International Conference on Communication Problem-Solving (ICCP)*, 233–236. https://doi.org/10.1109/ICCPS.2014.7062261

Duquesnoy, A. J., & Plomp, R. (1983). The effect of a hearing aid on the speech-reception threshold of hearing-impaired listeners in quiet and in noise. *The Journal of the Acoustical Society of America*, *73*(6), 2166–2173. https://doi.org/10.1121/1.389540

Dusan, S., & Rabiner, L. (2006). On the relation between maximum spectral transition positions and phone boundaries. *INTERSPEECH*, 645–648. Retrieved from http://www.isca-speech.org/archive/interspeech_2006/i06_1317.html

ETSI. (2007). *ETSI ES 202 050 V1.1.5 (2007-01) Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms* (Vol. 5). Retrieved from https://www.etsi.org/deliver/etsi_es/202000_202099/202050/01.01.05_60/es_202050v010105p.pdf

Fletcher, H. (1920). *Speech and hearing*. New York: D. Van Nostrand.

French, N. R., & Steinberg, J. C. (1947). Factors Governing the Intelligibility of Speech Sounds. *The Journal of the Acoustical Society of America1*, *19*, 90–119. https://doi.org/10.1121/1.1916407

Furui, S. (1986). On the role of spectral transition for speech perception. *The Journal of the Acoustical Society of America*, *80*(4), 1016–1025. https://doi.org/10.1121/1.393842

Ganchev, T. (2011). Contemporary Methods for Speech Parameterization. In *SpringerBriefs in Electrical and Computer Engineering*. https://doi.org/10.1007/978-1-4419-8447-0_1

Goldsworthy, R. L., & Greenberg, J. E. (2004). Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *The Journal of the Acoustical Society of America*, *116*, 3679–3689. https://doi.org/10.1121/1.1804628

Gopinath, B., Schneider, J., Hartley, D., Teber, E., McMahon, C. M., Leeder, S. R., & Mitchell, P. (2011). Incidence and Predictors of Hearing Aid Use and Ownership Among Older Adults With Hearing Loss. *Annals of Epidemiology*, *21*(7), 497–506. https://doi.org/10.1016/j.annepidem.2011.03.005

Gordon-Salant, S., & Fitzgibbons, P. J. (2004). Effects of stimulus and noise rate variability on speech perception by younger and older adults. *The Journal of the Acoustical Society of America*, *115*(4), 1808–1817. https://doi.org/10.1121/1.1645249

Gordon-Salant, S., Fitzgibbons, P. J., & Friedman, S. A. (2007). Recognition of Time-Compressed and Natural Speech With Selective Temporal Enhancements by Young and Elderly Listeners. *Journal of Speech, Language, and Hearing Research*, *50*(October), 1181–1194. https://doi.org/10.1044/1092-4388(2007/082)

Gordon-Salant, S., Fitzgibbons, P. J., & Yeni-Komshian, G. H. (2011). Auditory temporal processing and aging: implications for speech understanding of older people. *Audiology Research*, *1*(1S), 9–15. https://doi.org/10.4081/audiores.2011.e4

Gordon-Salant, S., & Friedman, S. A. (2011). Recognition of rapid speech by blind and sighted older adults. *Journal of Speech, Language, and Hearing Research*, *54*(2), 622–631. https://doi.org/10.1044/1092-4388(2010/10-0052)

Götzen, A. De, Bernardini, N., & Arfib, D. (2000). Traditional implementations of a phase-vocoder: the tricks of the trade. *COST-G6 Conference on Digital Audio Effects*, 1–7. Retrieved from http://www.cs.princeton.edu/courses/archive/spr09/cos325/Bernardini.pdf

Grofit, S., & Lavner, Y. (2008). Time-scale modification of audio signals using enhanced WSOLA with management of transients. *IEEE Transactions on Audio, Speech and Language Processing*, *16*(1), 106–115. https://doi.org/10.1109/TASL.2007.909444

Heute, U. (2006). Noise reduction. In E. Hänsler & G. Schmidt (Eds.), *Topics in Acoustic Echo and Noise Control. Signals and Communication Technology* (pp. 325–383). https://doi.org/10.1007/3-540-33213-8_9

Hindhede, A. L. (2011). Negotiating hearing disability and hearing disabled identities. *Health*, *16*(2), 169–185. https://doi.org/10.1177/1363459311403946

Hines, A., Skoglund, J., Kokaram, A. C., & Harte, N. (2015). ViSQOL: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, *2015*(13), 1–18. https://doi.org/10.1186/s13636-015-0054-9

Hines, A., Skoglund, J., Kokaram, A., & Harte, N. (2013). Robustness of Speech Quality Metrics to Background Noise and Network Degradations: Comparing VISQOL, PESQ and POLQA. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3697–3701. https://doi.org/10.1109/ICASSP.2013.6638348

Hirsch, H.-G., & Pearce, D. (2000). The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions. *ASR-2000*, 181–188. Retrieved from https://www.isca-speech.org/archive_open/archive_papers/asr2000/asr0_181.pdf

Hu, Y., & Loizou, P. C. (2007). Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication*, *49*(7–8), 588–601. https://doi.org/10.1016/j.specom.2006.12.006

IEEE Subcommittee on Subjective Measurements. (1969). IEEE Recommended Practice for Speech Quality Measurements. *IEEE Transactions on Audio and Electroacoustics*, *17*(3), 225–246. https://doi.org/10.1109/TAU.1969.1162058

Islam, M. T. (2020). Speech Enhancement Based on Probabilistic Geometric Spectral Subtraction. Retrieved April 5, 2020, from MATLAB Central File Exchange website: https://www.mathworks.com/matlabcentral/fileexchange/67292-speech-enhancement-based-on-probabilistic-geometric-spectral-subtraction

Islam, M. T., Shahnaz, C., Zhu, W.-P., & Ahmad, M. O. (2018). *Enhancement of Noisy Speech with Low Speech Distortion Based on Probabilistic Geometric Spectral Subtraction*. Retrieved from http://arxiv.org/abs/1802.05125

Iwasaki, S., Ocho, S., Nagura, M., & Hoshino, T. (2002). Contribution of speech rate to speech perception in multichannel cochlear implant users. *Annals of Otology, Rhinology and Laryngology*, *111*(8), 718–721. https://doi.org/10.1177/000348940211100811

Jassim, W. A., & Zilany, M. S. A. (2016). Speech quality assessment using 2D neurogram orthogonal moments. *Speech Communication*, *80*, 34–48. https://doi.org/10.1016/j.specom.2016.03.004

Jayan, A. R., Pandey, P. C., & Lehana, P. K. (2008). Automated detection of transition segments for intensity and time-scale modification for speech intelligibility enhancement. *IEEE International Conference on Signal Processing, Communications and Networking*, 63–68. https://doi.org/10.1109/ICSCN.2008.4447162

Jensen, J., & Taal, C. H. (2016). An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *24*(11), 2009–2022. https://doi.org/10.1109/TASLP.2016.2585878

Jenstad, L. M., & Souza, P. E. (2007). Temporal Envelope Changes of Compression and Speech Rate: Combined Effects on Recognition for Older Adults. *Journal of Speech, Language, and Hearing Research*, *50*, 1123–1138. https://doi.org/10.1044/1092-4388(2007/078)

Jin, S.-H., & Nelson, P. B. (2006). Speech perception in gated noise: The effects of temporal resolution. *The Journal of the Acoustical Society of America*, *119*(5), 3097–3108. https://doi.org/10.1121/1.2188688

Jin, S.-H., & Nelson, P. B. (2010). Interrupted speech perception: The effects of hearing sensitivity and frequency resolution. *The Journal of the Acoustical Society of America*, *128*(2), 881–889. https://doi.org/10.1121/1.3458851

Jørgensen, S., & 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*, 1475–1487. https://doi.org/10.1121/1.3621502

Kadiri, S. R., & Yegnanarayana, B. (2015). Analysis of singing voice for epoch extraction using Zero Frequency Filtering method. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4260–4264. https://doi.org/10.1109/ICASSP.2015.7178774

Kapilow, D., Stylianou, Y., & Schroeter, J. (1999). Detection of non-stationarity in speech signals and its application to time-scaling. *European Conference on Speech Communication and Technology (EUROSPEECH)*, 2307–2310. Retrieved from https://www.isca-speech.org/archive/archive_papers/eurospeech_1999/e99_2307.pdf

Kates, J. M., & Arehart, K. H. (2005). Coherence and the speech intelligibility index. *The Journal of the Acoustical Society of America*, *117*, 2224–2237. https://doi.org/10.1121/1.1862575

Kauppinen, I. (2002). Methods for detecting impulsive noise in speech and audio signals. *International Conference on Digital Signal Processing*, *July*, 967–970. https://doi.org/10.1109/ICDSP.2002.1028251

Kelly-Campbell, R. J., & Lessoway, K. (2015). Hearing aid and hearing assistance technology use in Aotearoa/New Zealand. *International Journal of Audiology*, *54*(5), 308–315. https://doi.org/10.3109/14992027.2014.979952

Kishon-Rabin, L., Avivi-Reich, M., & Roth, D. A.-E. (2013). Improved gap detection thresholds following auditory training: evidence of auditory plasticity in older adults. *American Journal of Audiology*, *22*, 343–346. https://doi.org/10.1044/1059-0889(2013/12-0084)

Kollmeier, B., Brand, T., & Meyer, B. (2008). Perception of Speech and Sound. In J. Benesty, M. M. Sondhi, & Y. Huang (Eds.), *Springer Handbook of Speech Processing* (pp. 61–82). Berlin, Heidelberg: Springer-Verlag.

Krueger, A., & Haeb-Umbach, R. (2010). Model-based feature enhancement for reverberant speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *18*(7), 1692–1707. https://doi.org/10.1109/TASL.2010.2049684

Kumar, P., Pradhan, B., Handa, D., & Sanju, H. K. (2016). Effect of Age on Time-Compressed Speech Perception and Speech Perception in Noise in Normal-Hearing Individuals. *Journal of Hearing Science*, *6*(1), 1–7. https://doi.org/10.17430/896978

Kupryjanow, A., & Czyzewski, A. (2012a). A Method of Real-Time Non-uniform Speech Stretching. In M. S. Obaidat, J. L. Sevillano, & J. Filipe (Eds.), *E-Business and Telecommunications* (pp. 362–373). Berlin, Heidelberg: Springer.

Kupryjanow, A., & Czyzewski, A. (2012b). Methods of Improving Speech Intelligibility for Listeners with Hearing Resolution Deficit. *Diagnostic Pathology*, *7*(129), 1–18. https://doi.org/10.1186/1746-1596-7-129

Lee, S., Kim, H. D., & Kim, H. S. (1997). Variable time-scale modification of speech using transient information. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *2*, 1319–1322. https://doi.org/10.1109/icassp.1997.596189

Lesica, N. A. (2018). Why Do Hearing Aids Fail to Restore Normal Auditory Perception? *Trends in Neurosciences*, *41*(4), 174–185. https://doi.org/10.1016/j.tins.2018.01.008

Lessa, A. H., & Costa, M. J. (2013). The impact of speech rate on sentence recognition by elderly individuals. *Brazilian Journal of Otorhinolaryngology*, *79*(6), 745–752. https://doi.org/10.5935/1808-8694.20130136

Leutnant, V., Krueger, A., & Haeb-Umbach, R. (2014). A new observation model in the logarithmic mel power spectral domain for the automatic recognition of noisy reverberant speech. *IEEE Transactions on Audio, Speech and Language Processing*, *22*(1), 95–109. https://doi.org/10.1109/TASLP.2013.2285480

Li, W., Wang, L., Zhou, Y., Bourlard, H., & Liao, Q. (2013). Robust log-energy estimation and

its dynamic change enhancement for in-car speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *21*(8), 1689–1698. https://doi.org/10.1109/TASL.2013.2260151

Lin, F. R., Yaffe, K., Xia, J., Xue, Q. L., Harris, T. B., Purchase-Helzner, E., … Simonsick, E. M. (2013). Hearing loss and cognitive decline in older adults. *JAMA Internal Medicine*, *173*(4), 293–299. https://doi.org/10.1001/jamainternmed.2013.1868

Loizou, P. C. (n.d.-a). MATLAB software. Retrieved April 5, 2020, from https://ecs.utdallas.edu/loizou/speech/software.htm

Loizou, P. C. (n.d.-b). NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms. Retrieved April 15, 2020, from https://ecs.utdallas.edu/loizou/speech/noizeus/

Loughrey, D. G., Kelly, M. E., Kelley, G. A., Brennan, S., & Lawlor, B. A. (2018). Association of Age-Related Hearing Loss With Cognitive Function, Cognitive Impairment, and Dementia. *JAMA Otolaryngology–Head & Neck Surgery*, *144*(2), 115–126. https://doi.org/10.1001/jamaoto.2017.2513

Lu, Y., & Loizou, P. C. (2008). A geometric approach to spectral subtraction. *Speech Communication*, *50*(6), 453–466. https://doi.org/10.1016/j.specom.2008.01.003

Ma, J., & Loizou, P. C. (2011). SNR loss : A new objective measure for predicting the intelligibility of noise-suppressed speech. *Speech Communication*, *53*(3), 340–354. https://doi.org/10.1016/j.specom.2010.10.005

Ma, Y., & Nishihara, A. (2014). A modified Wiener filtering method combined with wavelet thresholding multitaper spectrum for speech enhancement. *EURASIP Journal on Audio, Speech, and Music Processing*, *2014*(32), 1–11. https://doi.org/10.1186/s13636-014-0032-7

Maharani, A., Dawes, P., Nazroo, J., Tampubolon, G., & Pendleton, N. (2018). Longitudinal Relationship Between Hearing Aid Use and Cognitive Function in Older Americans. *Journal of the American Geriatrics Society*, *66*(6), 1130–1136. https://doi.org/10.1111/jgs.15363

Mathers, C. D., & Loncar, D. (2006). Projections of Global Mortality and Burden of Disease from 2002 to 2030. *PLoS Medicine*, *3*(11), 2011–2030. https://doi.org/10.1371/journal.pmed.0030442

Mick, P., Kawachi, I., & Lin, F. R. (2014). The association between hearing loss and social isolation in older adults. *Otolaryngology - Head and Neck Surgery (United States)*, *150*(3),

378–384. https://doi.org/10.1177/0194599813518021

Morgan, N., & Fosler-Lussier, E. (1998). Combining multiple estimators of speaking rate. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 729–732. https://doi.org/10.1109/ICASSP.1998.675368

Morgan, N., Fosler, E., & Mirghafori, N. (1997). Speech Recognition Using On-line Estimation of Speaking Rate. *European Conference on Speech Communication and Technology (EUROSPEECH)*, 2079–2082. Retrieved from https://www.isca-speech.org/archive/archive_papers/eurospeech_1997/e97_2079.pdf

Moulines, E., & Charpentier, F. (1990). Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. *Speech Communication*, *9*(5–6), 453–467. https://doi.org/10.1016/0167-6393(90)90021-Z

Moulines, E., & Laroche, J. (1995). Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, *16*(2), 175–205. https://doi.org/10.1016/0167-6393(94)00054-E

Murty, K. S. R., & Yegnanarayana, B. (2008). Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech and Language Processing*, *16*(8), 1602–1613. https://doi.org/10.1109/TASL.2008.2004526

Nejime, Y., & Moore, B. C. J. (1998). Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss. *The Journal of the Acoustical Society of America*, *103*(1), 572–576. https://doi.org/10.1121/1.421123

Ni, J., Shiga, Y., & Kawai, H. (2016). Using Zero-Frequency Resonator to Extract Multilingual Intonation Structure. *INTERSPEECH*, 1522–1526. https://doi.org/10.21437/Interspeech.2016-1607

Nicholson, W. K. (2006). *Linear Algebra with Applications* (5th ed.). Canada: McGraw-Hill Ryerson Limited.

NOIZEUS speech corpus. (2018). Retrieved April 15, 2020, from https://github.com/chmodsss/noizeus_corpora

Oehmen, R., Kirsner, K., & Fay, N. (2010). Reliability of the Manual Segmentation of Pauses in Natural Speech. In H. Loftsson, E. Rögnvaldsson, & S. Helgadóttir (Eds.), *Advances in Natural Language Processing. NLP 2010. Lecture Note in Computer Science, vol 6233* (pp. 263–268). https://doi.org/10.1007/978-3-642-14770-8_30

Patil, U. G., & Shirbahadurkar, S. D. (2018). Performance analysis of SS based speech enhancement algorithms for ASR with Non-stationary Noisy Database-NOIZEUS. *International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*, 636–641. https://doi.org/10.1109/I-SMAC.2018.8653675

Payton, K. L., Uchanski, R. M., & Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *Journal of the Acoustical Society of America*, *95*(3), 1581–1592. https://doi.org/10.1121/1.408545

Peters, R. W., Moore, B. C. J., & Baer, T. (1998). Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. *The Journal of the Acoustical Society of America*, *103*(1), 577–587. https://doi.org/10.1121/1.421128

Pleis, J. R., Ward, B. W., & Lucas, J. W. (2010). Summary Health Statistics for U . S . Adults: National Health Interview Survey, 2009. *National Center for Health Statistics. Vital Health Statistics*, *10*(249), 1–217.

Portney, L. G., & Watkins, M. P. (2015). *Foundations of Clinical Research: Applications to Practice* (3rd ed.). USA: F. A. Davis Company.

Quatieri, T. F., & McAulay, R. J. (1992). Shape Invariant Time-Scale and Pitch Modification of Speech. *IEEE Transactions on Signal Processing*, *40*(3), 497–510. https://doi.org/10.1109/78.120793

Rudresh, S., Vasisht, A., Vijayan, K., & Seelamantula, C. S. (2018). Epoch-Synchronous Overlap-Add (ESOLA) for Time- and Pitch-Scale Modification of Speech Signals. *ArXiv Preprint*, 1–10. Retrieved from http://arxiv.org/abs/1801.06492

Salonen, J., Johansson, R., Karjalainen, S., Vahlberg, T., Jero, J. P., & Isoaho, R. (2013). Hearing aid compliance in the elderly. *B-ENT*, *9*(1), 23–28.

Sao, A. K., & Yegnanarayana, B. (2012). Edge extraction using zero-frequency resonator. *Signal, Image and Video Processing*, *6*, 287–300. https://doi.org/10.1007/s11760-011-0233-9

Schafer, R. W. (2008). Homomorphic Systems and Cepstrum Analysis of Speech. In J. Benesty, M. M. Sondhi, & Y. Huang (Eds.), *Springer Handbook of Speech Processing* (pp. 161–180). Springer-Verlag.

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in

auditory scene analysis. *Trends in Neurosciences*, *34*(3), 114–123. https://doi.org/10.1016/j.tins.2010.11.002

Shi, Y., Peng, K. A., Chen, B., Gong, Y., Chen, J., Li, Y., & Fu, Q. J. (2018). Interaction between speech variations and background noise on speech intelligibility by Mandarin-speaking cochlear implant patients. *Speech Communication*, *104*, 89–94. https://doi.org/10.1016/j.specom.2018.09.007

Song, J. H., Skoe, E., Banai, K., & Kraus, N. (2012). Training to Improve Hearing Speech in Noise: Biological Mechanisms. *Cerebral Cortex*, *22*(5), 1180–1190. https://doi.org/10.1093/cercor/bhr196

Sound Zone Tools: Signal Processing Tools for MATLAB. (2017). Retrieved April 15, 2020, from https://github.com/JacobD10/SoundZone_Tools

Sreenivasa Rao, K., & Vuppala, A. K. (2013). Non-uniform time scale modification using instants of significant excitation and vowel onset points. *Speech Communication*, *55*(6), 745–756. https://doi.org/10.1016/j.specom.2013.03.002

Sreenivasa Rao, K., & Yegnanarayana, B. (2006). Prosody modification using instants of significant excitation. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(3), 972–980. https://doi.org/10.1109/TSA.2005.858051

Stecker, G. C., Bowman, G. A., Yund, W., Herron, T. J., Roup, C. M., & Woods, D. L. (2006). Perceptual training improves syllable identification in new and experienced hearing aid users. *Journal of Rehabilitation Research and Development*, *43*(4), 537–551. https://doi.org/10.1682/JRRD.2005.11.0171

Steeneken, H. J. M., & Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, *67*, 318–326. https://doi.org/10.1121/1.384464

Stollman, M. H. P., & Kapteyn, T. S. (1994). Effect of time scale modification of speech on the speech recognition threshold in noise for elderly listeners. *International Journal of Audiology*, *33*(5), 280–290. https://doi.org/10.3109/00206099409071888

Stollman, M. H. P., Kapteyn, T. S., & Sleeswijk, B. W. (1994). Effect of time-scale modification of speech on the speech recognition threshold in noise for hearing-impaired and language-impaired children. *Scandinavian Audiology*, *23*(1), 39–46. https://doi.org/10.3109/01050399409047484

Sweetow, R. W., & Sabes, J. H. (2006). The Need for and Development of an Adaptive Listening and Communication Enchancement (LACE^{TM}) Program. *Journal of the American Academy of Audiology*, *17*(8), 538–558. https://doi.org/10.3766/jaaa.17.8.2

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4214–4217. https://doi.org/10.1109/ICASSP.2010.5495701

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 2125–2136. https://doi.org/10.1109/TASL.2011.2114881

Tremblay, K. L. (2015). The Ear – Brain Connection: Older Ears and Older Brains. *American Journal of Audiology*, *24*, 117–121. https://doi.org/10.1044/2015_AJA-14-0068

Tremblay, K. L., & Kraus, N. (2002). Auditory training induces asymmetrical changes in cortical neural activity. *Journal of Speech, Language, and Hearing Research*, *45*(3), 564–572. https://doi.org/10.1044/1092-4388(2002/045)

Tremblay, K. L., Pinto, A., Fischer, M. E., Klein, B. E. K., Klein, R., Levy, S., … Cruickshanks, K. J. (2015). Self-Reported Hearing Difficulties Among Adults With Normal Audiograms: The Beaver Dam Offspring Study. *Ear and Hearing*, *36*(6), e290–e299. https://doi.org/10.1097/AUD.0000000000000195

Tremblay, K. L., Piskosz, M., & Souza, P. (2003). Effects of age and age-related hearing loss on the neural representation of speech cues. *Clinical Neurophysiology*, *114*(7), 1332–1343. https://doi.org/10.1016/S1338-2457(03)00114-7

Turan, M. A. T., & Erzin, E. (2015). Synchronous Overlap and Add of Spectra for Enhancement of Excitation in Artificial Bandwidth Extension of Speech. *INTERSPEECH*, 2588–2592. Dresden, Germany.

Valbret, H., Moulines, E., & Tubach, J. P. (1992). Voice tranformation using PSOLA technique. *Speech Communication*, *11*(2–3), 175–187. https://doi.org/10.1016/0167-6393(92)90012-V

Verhelst, W., & Roelands, M. (1993). An Overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *2*, 554–557.

https://doi.org/10.1109/icassp.1993.319366

Xia, J., Xu, B., Pentony, S., Xu, J., & Swaminathan, J. (2018). Effects of reverberation and noise on speech intelligibility in normal-hearing and aided hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *143*(3), 1523–1533. https://doi.org/10.1121/1.5026788

Yegnanarayana, B., & Gangashetty, S. V. (2011). Epoch-based analysis of speech signals. *Sadhana - Academy Proceedings in Engineering Sciences*, *36*(5), 651–697. https://doi.org/10.1007/s12046-011-0046-0

Yegnanarayana, B., Prasanna, S. R. M., & Guruprasad, S. (2011). Study of Robustness of Zero Frequency Resonator Method for Extraction of Fundamental Frequency. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5392–5395. https://doi.org/10.1109/ICASSP.2011.5947577

Yost, W. A. (2000). *Fundamentals of Hearing: An Introduction* (4th ed.). San Diego, CA, USA: Elsevier Science.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., … Woodland, P. (2009). *The HTK Book 3.4* (Vol. 32, p. 680). Vol. 32, p. 680. Retrieved from http://www.inf.u-szeged.hu/~tothl/speech/htkbook.pdf

Zheng, F., Zhang, G., & Song, Z. (2001). Comparison of different implementations of MFCC. *Journal of Computer Science and Technology*, *16*(6), 582–589. https://doi.org/10.1007/BF02943243

Appendix A

Table A: Correlation between STOI and MFCC-SRBIM calculated using different combinations of threshold and window length ($p < 0.001$). Threshold 1.0 (highlighted row) showed the most negative correlations

	Window length (ms)									
Thres-hold	30	50	70	90	110	130	150	170	190	210
0.2	0.748 ±0.005	0.777 ±0.003	0.759 ±0.004	0.727 ±0.004	0.742 ±0.004	0.756 ±0.003	0.748 ±0.004	0.747 ±0.003	0.747 ±0.003	0.735 ±0.004
0.4	0.532 ±0.007	0.672 ±0.005	0.694 ±0.004	0.678 ±0.005	0.663 ±0.005	0.653 ±0.006	0.656 ±0.006	0.638 ±0.007	0.622 ±0.007	0.607 ±0.007
0.6	-0.147 ±0.009	0.187 ±0.009	0.19 ±0.01	0.15 ±0.01	0.155 ±0.008	0.130 ±0.007	*0.028 ±0.008*	*0.008 ±0.009*	*-0.043 ±0.009*	*-0.069 ±0.009*
0.8	-0.545 ±0.008	-0.405 ±0.009	-0.460 ±0.007	-0.514 ±0.007	-0.470 ±0.008	-0.480 ±0.007	-0.487 ±0.008	-0.472 ±0.008	-0.468 ±0.008	-0.490 ±0.008
1.0	-0.665 ±0.006	-0.610 ±0.006	-0.668 ±0.005	-0.689 ±0.005	-0.640 ±0.006	-0.632 ±0.006	-0.638 ±0.005	-0.639 ±0.005	-0.617 ±0.006	-0.617 ±0.007
1.2	-0.597 ±0.006	-0.425 ±0.009	-0.487 ±0.009	-0.503 ±0.007	-0.491 ±0.007	-0.526 ±0.007	-0.559 ±0.005	-0.551 ±0.009	-0.548 ±0.007	-0.571 ±0.007
1.4	-0.450 ±0.009	-0.16 ±0.01	-0.16 ±0.01	-0.21 ±0.01	-0.275 ±0.009	-0.280 ±0.008	-0.323 ±0.006	-0.361 ±0.006	-0.345 ±0.008	-0.368 ±0.008
1.6	-0.167 ±0.008	*0.103 ±0.008*	0.108 ±0.008	*0.065 ±0.009*	*0.04 ±0.01*	*0.074 ±0.009*	*0.029 ±0.009*	*-0.016 ±0.008*	*-0.045 ±0.009*	*-0.057 ±0.009*
1.8	0.117 ±0.009	0.346 ±0.008	0.354 ±0.008	0.285 ±0.008	0.310 ±0.010	0.333 ±0.009	0.31 ±0.01	0.260 ±0.009	0.229 ±0.008	0.225 ±0.008
2.0	0.365 ±0.007	0.561 ±0.006	0.565 ±0.007	0.482 ±0.007	0.434 ±0.007	0.490 ±0.007	0.470 ±0.009	0.452 ±0.008	0.418 ±0.007	0.403 ±0.007
2.2	0.484 ±0.006	0.651 ±0.004	0.619 ±0.005	0.590 ±0.006	0.555 ±0.008	0.56 ±0.01	0.52 ±0.02	0.522 ±0.009	0.537 ±0.007	0.543 ±0.008
2.4	0.560 ±0.006	0.664 ±0.008	0.667 ±0.007	0.644 ±0.007	0.642 ±0.005	0.602 ±0.007	0.57 ±0.02	0.627 ±0.007	0.622 ±0.007	0.642 ±0.005
2.6	0.619 ±0.006	0.695 ±0.007	0.704 ±0.006	0.670 ±0.006	0.663 ±0.005	0.656 ±0.006	0.65 ±0.01	0.61 ±0.01	0.680 ±0.006	0.681 ±0.006
2.8	0.660 ±0.005	0.701 ±0.005	0.723 ±0.006	0.695 ±0.005	0.657 ±0.005	0.678 ±0.005	0.697 ±0.008	0.671 ±0.009	0.694 ±0.007	0.685 ±0.007
3.0	0.698 ±0.004	0.718 ±0.005	0.725 ±0.005	0.705 ±0.005	0.681 ±0.005	0.686 ±0.004	0.691 ±0.007	0.700 ±0.007	0.695 ±0.007	0.696 ±0.007

Appendix B

Table B: Correlation between STOI and LMPSC-SRBIM calculated using different combinations of threshold and window length ($p < 0.001$). Threshold 1.0 (highlighted row) showed the most negative correlations

	Window length (ms)									
Thres-hold	30	50	70	90	110	130	150	170	190	210
0.2	0.784 ±0.003	0.778 ±0.003	0.692 ±0.006	0.645 ±0.005	0.623 ±0.004	0.662 ±0.004	0.658 ±0.005	0.656 ±0.005	0.672 ±0.004	0.667 ±0.004
0.4	0.661 ±0.007	0.772 ±0.003	0.760 ±0.005	0.739 ±0.004	0.754 ±0.004	0.763 ±0.004	0.742 ±0.004	0.744 ±0.004	0.723 ±0.004	0.715 ±0.006
0.6	0.083 ±0.008	0.405 ±0.008	0.403 ±0.009	0.406 ±0.009	0.374 ±0.008	0.359 ±0.009	0.28 ±0.01	0.25 ±0.01	0.18 ±0.01	0.131 ±0.009
0.8	-0.557 ±0.007	-0.375 ±0.009	-0.394 ±0.008	-0.42 ±0.01	-0.451 ±0.007	-0.461 ±0.007	-0.431 ±0.009	-0.438 ±0.009	-0.479 ±0.008	-0.473 ±0.009
1.0	-0.606 ±0.007	-0.596 ±0.008	-0.656 ±0.007	-0.655 ±0.008	-0.675 ±0.004	-0.648 ±0.006	-0.676 ±0.005	-0.645 ±0.006	-0.652 ±0.007	-0.641 ±0.007
1.2	-0.549 ±0.006	-0.410 ±0.008	-0.444 ±0.009	-0.469 ±0.008	-0.448 ±0.008	-0.480 ±0.008	-0.513 ±0.007	-0.504 ±0.007	-0.493 ±0.008	-0.479 ±0.008
1.4	-0.28 ±0.01	0.059 ±0.007	0.001 ±0.009	-0.050 ±0.009	-0.131 ±0.008	-0.078 ±0.008	-0.121 ±0.008	-0.177 ±0.008	-0.232 ±0.007	-0.234 ±0.007
1.6	0.093 ±0.007	0.398 ±0.007	0.401 ±0.007	0.307 ±0.008	0.308 ±0.008	0.340 ±0.009	0.329 ±0.008	0.269 ±0.008	0.253 ±0.009	0.210 ±0.009
1.8	0.387 ±0.007	0.561 ±0.007	0.554 ±0.007	0.484 ±0.007	0.448 ±0.009	0.517 ±0.010	0.543 ±0.008	0.498 ±0.008	0.461 ±0.008	0.467 ±0.007
2.0	0.570 ±0.006	0.679 ±0.005	0.646 ±0.006	0.571 ±0.006	0.569 ±0.008	0.588 ±0.009	0.60 ±0.01	0.615 ±0.008	0.623 ±0.007	0.606 ±0.006
2.2	0.634 ±0.007	0.701 ±0.006	0.659 ±0.007	0.635 ±0.008	0.638 ±0.007	0.666 ±0.006	0.650 ±0.007	0.662 ±0.006	0.679 ±0.006	0.661 ±0.006
2.4	0.692 ±0.007	0.731 ±0.005	0.710 ±0.005	0.712 ±0.006	0.698 ±0.006	0.686 ±0.009	0.66 ±0.01	0.700 ±0.007	0.724 ±0.007	0.731 ±0.006
2.6	0.703 ±0.005	0.757 ±0.004	0.734 ±0.005	0.720 ±0.005	0.710 ±0.006	0.693 ±0.008	0.68 ±0.01	0.680 ±0.009	0.731 ±0.007	0.754 ±0.005
2.8	0.738 ±0.004	0.779 ±0.004	0.753 ±0.005	0.735 ±0.005	0.723 ±0.005	0.697 ±0.007	0.701 ±0.009	0.675 ±0.009	0.722 ±0.007	0.744 ±0.006
3.0	0.744 ±0.005	0.767 ±0.004	0.747 ±0.005	0.736 ±0.005	0.712 ±0.006	0.712 ±0.006	0.710 ±0.008	0.68 ±0.01	0.727 ±0.007	0.747 ±0.006

Appendix C

Table C: Correlation between SNR Loss and MFCC-SRBIM calculated using different combinations of threshold and window length ($p < 0.001$). Threshold 1.0 (highlighted row) showed the most positive correlations

	Window length (ms)									
Thres-hold	30	50	70	90	110	130	150	170	190	210
0.2	-0.631 ±0.008	-0.667 ±0.005	-0.686 ±0.004	-0.718 ±0.004	-0.741 ±0.003	-0.731 ±0.003	-0.723 ±0.004	-0.712 ±0.004	-0.710 ±0.003	-0.693 ±0.004
0.4	-0.440 ±0.005	-0.538 ±0.004	-0.581 ±0.004	-0.588 ±0.004	-0.566 ±0.004	-0.530 ±0.005	-0.527 ±0.004	-0.525 ±0.004	-0.523 ±0.004	-0.508 ±0.004
0.6	0.150 ±0.007	-0.115 ±0.006	-0.118 ±0.006	*-0.097 ±0.008*	*-0.090 ±0.008*	*-0.055 ±0.007*	*-0.015 ±0.006*	*0.005 ±0.006*	*0.041 ±0.007*	*0.071 ±0.007*
0.8	0.45 ±0.01	0.353 ±0.007	0.386 ±0.006	0.417 ±0.006	0.388 ±0.006	0.407 ±0.005	0.416 ±0.006	0.414 ±0.006	0.420 ±0.005	0.420 ±0.005
1.0	0.511 ±0.006	0.474 ±0.006	0.526 ±0.004	0.537 ±0.005	0.533 ±0.005	0.505 ±0.005	0.522 ±0.004	0.514 ±0.004	0.485 ±0.004	0.478 ±0.005
1.2	0.490 ±0.005	0.350 ±0.004	0.397 ±0.005	0.423 ±0.005	0.406 ±0.006	0.424 ±0.005	0.412 ±0.004	0.397 ±0.005	0.409 ±0.004	0.416 ±0.005
1.4	0.360 ±0.005	0.115 ±0.005	0.110 ±0.006	0.178 ±0.007	0.212 ±0.007	0.217 ±0.006	0.213 ±0.005	0.224 ±0.005	0.228 ±0.006	0.239 ±0.006
1.6	0.135 ±0.006	-0.115 ±0.006	-0.129 ±0.006	*-0.081 ±0.007*	*-0.072 ±0.007*	*-0.087 ±0.007*	*-0.042 ±0.008*	*-0.031 ±0.007*	*-0.003 ±0.007*	*0.019 ±0.006*
1.8	*-0.102 ±0.007*	-0.293 ±0.006	-0.335 ±0.006	-0.266 ±0.007	-0.264 ±0.008	-0.271 ±0.006	-0.239 ±0.008	-0.243 ±0.006	-0.217 ±0.006	-0.217 ±0.006
2.0	-0.307 ±0.005	-0.447 ±0.004	-0.459 ±0.006	-0.407 ±0.007	-0.380 ±0.006	-0.414 ±0.006	-0.386 ±0.008	-0.386 ±0.005	-0.357 ±0.006	-0.356 ±0.006
2.2	-0.407 ±0.005	-0.532 ±0.003	-0.533 ±0.003	-0.501 ±0.006	-0.477 ±0.008	-0.456 ±0.009	-0.41 ±0.01	-0.422 ±0.008	-0.435 ±0.006	-0.435 ±0.008
2.4	-0.476 ±0.004	-0.556 ±0.005	-0.580 ±0.004	-0.545 ±0.007	-0.557 ±0.005	-0.502 ±0.006	-0.45 ±0.02	-0.494 ±0.007	-0.503 ±0.006	-0.518 ±0.005
2.6	-0.525 ±0.004	-0.572 ±0.004	-0.591 ±0.005	-0.556 ±0.006	-0.561 ±0.004	-0.535 ±0.005	-0.50 ±0.01	-0.47 ±0.01	-0.523 ±0.005	-0.541 ±0.006
2.8	-0.560 ±0.004	-0.582 ±0.004	-0.593 ±0.004	-0.568 ±0.005	-0.556 ±0.004	-0.544 ±0.005	-0.544 ±0.008	-0.520 ±0.009	-0.527 ±0.007	-0.537 ±0.005
3.0	-0.566 ±0.006	-0.583 ±0.004	-0.581 ±0.004	-0.560 ±0.005	-0.550 ±0.005	-0.549 ±0.005	-0.547 ±0.006	-0.556 ±0.005	-0.534 ±0.006	-0.537 ±0.007

Appendix D

Table D: Correlation between SNR Loss and LMPSC-SRBIM calculated using different combinations of threshold and window Length ($p < 0.001$). Threshold 1.0 (highlighted row) showed the most positive correlations

	Window length (ms)									
Thres-hold	30	50	70	90	110	130	150	170	190	210
0.2	-0.708 ±0.004	-0.732 ±0.004	-0.701 ±0.005	-0.748 ±0.003	-0.712 ±0.005	-0.760 ±0.003	-0.741 ±0.004	-0.743 ±0.006	-0.752 ±0.005	-0.746 ±0.005
0.4	-0.556 ±0.005	-0.648 ±0.003	-0.657 ±0.007	-0.676 ±0.004	-0.677 ±0.003	-0.669 ±0.003	-0.644 ±0.002	-0.634 ±0.003	-0.623 ±0.003	-0.620 ±0.004
0.6	*-0.059 ±0.005*	-0.324 ±0.005	-0.336 ±0.005	-0.316 ±0.006	-0.272 ±0.006	-0.247 ±0.008	-0.206 ±0.006	-0.185 ±0.006	-0.154 ±0.006	*-0.101 ±0.006*
0.8	0.462 ±0.007	0.321 ±0.006	0.352 ±0.005	0.367 ±0.007	0.394 ±0.004	0.389 ±0.005	0.387 ±0.006	0.406 ±0.005	0.415 ±0.005	0.410 ±0.005
1.0	0.506 ±0.006	0.474 ±0.006	0.524 ±0.005	0.531 ±0.005	0.550 ±0.005	0.511 ±0.004	0.529 ±0.003	0.522 ±0.005	0.520 ±0.005	0.507 ±0.005
1.2	0.438 ±0.005	0.318 ±0.005	0.356 ±0.005	0.370 ±0.005	0.374 ±0.006	0.365 ±0.006	0.372 ±0.005	0.388 ±0.005	0.392 ±0.006	0.383 ±0.006
1.4	0.249 ±0.005	*-0.023 ±0.006*	*0.008 ±0.006*	*0.048 ±0.006*	*0.075 ±0.007*	*0.048 ±0.007*	*0.099 ±0.007*	*0.098 ±0.006*	0.118 ±0.006	0.148 ±0.007
1.6	*-0.089 ±0.006*	-0.318 ±0.007	-0.356 ±0.005	-0.281 ±0.006	-0.293 ±0.006	-0.307 ±0.005	-0.274 ±0.006	-0.239 ±0.004	-0.202 ±0.005	-0.184 ±0.006
1.8	-0.327 ±0.007	-0.463 ±0.006	-0.466 ±0.007	-0.440 ±0.006	-0.402 ±0.009	-0.41 ±0.01	-0.443 ±0.006	-0.409 ±0.005	-0.392 ±0.006	-0.380 ±0.005
2.0	-0.480 ±0.005	-0.552 ±0.005	-0.560 ±0.005	-0.507 ±0.007	-0.493 ±0.009	-0.470 ±0.008	-0.472 ±0.009	-0.475 ±0.006	-0.504 ±0.005	-0.494 ±0.005
2.2	-0.532 ±0.006	-0.569 ±0.005	-0.569 ±0.006	-0.529 ±0.008	-0.542 ±0.008	-0.543 ±0.006	-0.524 ±0.006	-0.514 ±0.006	-0.532 ±0.006	-0.533 ±0.006
2.4	-0.579 ±0.005	-0.614 ±0.003	-0.603 ±0.005	-0.594 ±0.006	-0.587 ±0.006	-0.558 ±0.008	-0.53 ±0.01	-0.556 ±0.007	-0.557 ±0.007	-0.579 ±0.005
2.6	-0.594 ±0.004	-0.620 ±0.003	-0.616 ±0.004	-0.614 ±0.005	-0.580 ±0.005	-0.568 ±0.008	-0.55 ±0.01	-0.539 ±0.009	-0.569 ±0.007	-0.597 ±0.005
2.8	-0.628 ±0.003	-0.636 ±0.003	-0.625 ±0.004	-0.603 ±0.004	-0.582 ±0.004	-0.550 ±0.008	-0.553 ±0.009	-0.536 ±0.009	-0.577 ±0.006	-0.600 ±0.004
3.0	-0.630 ±0.003	-0.626 ±0.004	-0.608 ±0.007	-0.593 ±0.007	-0.577 ±0.006	-0.554 ±0.006	-0.567 ±0.007	-0.53 ±0.01	-0.572 ±0.008	-0.598 ±0.006