COPPE
UFRJ
Instituto Alberto Luiz Coimbra de
Pós-Graduação e Pesquisa de Engenharia

ON THE ENHANCEMENT OF SPEECH DEREVERBERATION ALGORITHMS USING MULTIPLE OBJECTIVE ASSESSMENT MEASURES

Rafael Zambrano López

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientadores: Sergio Lima Netto
Thiago de Moura Prego

Rio de Janeiro
Agosto de 2016

ON THE ENHANCEMENT OF SPEECH DEREVERBERATION ALGORITHMS USING MULTIPLE OBJECTIVE ASSESSMENT MEASURES

Rafael Zambrano López

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. Sergio Lima Netto, Ph.D.

Prof. Thiago de Moura Prego, D.Sc.

Prof. Luiz Wagner Pereira Biscainho, D.Sc.

Prof. Lisandro Lovisolo, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
AGOSTO DE 2016

López, Rafael Zambrano

On the Enhancement of Speech Dereverberation Algorithms Using Multiple Objective Assessment Measures/Rafael Zambrano López. – Rio de Janeiro: UFRJ/COPPE, 2016.

XVIII, 86 p.: il.; 29, 7cm.

Orientadores: Sergio Lima Netto
Thiago de Moura Prego

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2016.

Referências Bibliográficas: p. 79 – 85.

1. Speech Dereverberation. 2. Assessment Measures. 3. Multi-Objective Optimization. I. Lima Netto, Sergio *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

To Clarice

Acknowledgements

First, I want to take the opportunity in these lines to express thanks and gratitude to my advisor, Sergio Lima Netto, for the guidance, support and encouragement, both personal and professional. His influence has been invaluable in the last years.

I thank the teachers Amaro Azevedo de Lima and Thiago de Moura Prego (also advisor of this work), for their advice and dedication during the preparation of this work.

I thank the CAPES agency for the scholarship received during the master.

Thanks to teachers Luiz Wagner Pereira Biscainho and Lisandro Lovisolo for agreeing to take part in the evaluation team. Thanks also to all the members of the Signal Multimedia and Telecommunications Laboratory (SMT), the best working group I have ever met.

I sincerely thank Scott Wisdom and Dayana Ribas González for sharing their dereverberation algorithms, which were essential in the practical implementation of this work.

Last but not least, I thank my family and friends for their continued support.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

ACERCA DO APERFEIÇOAMENTO DE ALGORITMOS DE DESREVERBERAÇÃO USANDO MÚLTIPLAS MÉTRICAS OBJETIVAS DE AVALIAÇÃO

Rafael Zambrano López

Agosto/2016

Orientadores: Sergio Lima Netto
Thiago de Moura Prego

Programa: Engenharia Elétrica

Este trabalho apresenta uma metodologia para melhorar o desempenho de algoritmos de desreverberação. A reverberação é um fenômeno que afeta negativamente o desempenho de inúmeras técnicas e sistemas de processamento de sinais de fala, tais como sistemas de reconhecimento e dispositivos de apoio a deficientes auditivos. A utilização de algoritmos de desreverberação é portanto necessária para melhorar a qualidade e inteligibilidade de sinais de fala degradados pela reverberação. Para avaliar o desempenho desses algoritmos, diversas métricas de avaliação objetivas têm sido desenvolvidas nos últimos tempos. A metodologia proposta neste trabalho será feita através da otimização de várias dessas métricas simultaneamente, com o objetivo de obter uma nova configuração do algoritmo de desreverberação que supere, em termos de qualidade e inteligibilidade dos sinais de fala processados, a sua configuração original.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

ON THE ENHANCEMENT OF SPEECH DEREVERBERATION ALGORITHMS USING MULTIPLE OBJECTIVE ASSESSMENT MEASURES

Rafael Zambrano López

August/2016

Advisors: Sergio Lima Netto
Thiago de Moura Prego

Department: Electrical Engineering

This work provides a methodology for improving the performance of dereverberation algorithms. Reverberation is a phenomenon that adversely affects the performance of numerous speech signal processing techniques and systems, such as speech recognition and hearing aid systems. The use of dereverberation algorithms is thereby necessary to improve the quality and intelligibility of speech signals degraded by reverberation. To evaluate the performance of these algorithms, numerous assessment objective measures have been developed in recent times. The methodology proposed in this work will be carried out by means of the optimization of several of these measures simultaneously, in order to obtain a new configuration of the dereverberation algorithm which outperforms, in terms of quality and intelligibility of the processed speech signals, its original configuration.

Contents

List of Figures xi

List of Tables xiv

List of Symbols xvi

List of Abbreviations xvii

1 Introduction 1

1.1 Motivation and Work Proposal 2

1.2 Organization of the Thesis 2

2 Fundamentals of Reverberation 3

2.1 Reverberation and its Effects 3

2.1.1 Effects of Reverberation in the Frequency Domain 5

2.1.2 Effects of Reverberation on Automatic Speech Recognition . . 7

2.2 Acoustic Impulse Responses 8

2.2.1 Reverberation Time 10

2.2.2 Direct-to-Reverberant Ratio 10

2.2.3 Room Spectral Variance 11

2.3 Dereverberation Systems 11

3 Evaluating the Performance of Speech Dereverberation Algorithms 13

3.1 Quality and Intelligibility of Speech 13

3.2 Subjective Measures 14

3.3 Objective Measures 15

3.3.1 Reverberation-Oriented Quality Measures 15

3.3.1.1 Allen's Measure 15

3.3.1.2 QAreverb 16

3.3.1.3 Speech to Reverberation Modulation Energy Ratio . 16

3.3.1.4 Other Reverberation-Oriented Measures 17

3.3.2 General Purpose Quality Measures 17

3.3.2.1 Perceptual Evaluation of Speech Quality 17
3.3.2.2 Other General Purpose Quality Measures 18
3.3.3 Speech Intelligibility Measures 19
3.3.3.1 Normalized-Covariance Measure 19
3.3.3.2 Word Error Rate 20
3.4 Conclusion 20

4 Proposed Methodology for Enhancing Dereverberation Algorithms 21
4.1 Multi-Objective Optimization and Pareto-Optimal Solutions 21
4.2 Proposed Method 24
4.3 Practical Considerations 26
4.3.1 Reverberant Speech Databases 27
4.3.1.1 Training Set 27
4.3.2 Speech Recognition System 28
4.3.3 Objective Assessment Measures 28
4.4 Conclusion 29

5 Enhancement of Dereverberation Algorithm A1 30
5.1 A1 description 30
5.2 Enhancement Process 32
5.2.1 Choice of the Parameters 32
5.2.2 Choice of Objective-Assessment Measures 33
5.2.3 Choice of the Optimal Operating Point 33
5.2.3.1 Experiment 1 36
5.2.3.2 Experiment 2 39
5.2.3.3 Experiment 3 39
5.3 Overall Results 41
5.4 Conclusion 47

6 Enhancement of Dereverberation Algorithm A2 48
6.1 A2 Description 48
6.1.1 Minimum Mean Square Error Log Spectral Amplitude estimator 49
6.1.1.1 *A priori* SIR Estimator 51
6.1.1.2 Estimator for the Ambient Noise Spectral Variance . 52
6.1.1.3 Estimator for the Late Reverberant Spectral Variance 52
6.2 Enhancement Process 53
6.2.1 Choice of the parameters 53
6.2.2 Choice of the Optimal Operating Point 55
6.3 Overall Results 57
6.4 Conclusion 63

7 Enhancement of Dereverberation Algorithm A3 64

7.1 A3 Description 64

7.2 Enhancement Process 65

7.2.1 Choice of the Parameters 65

7.2.2 Choice of the Optimal Operating Point 66

7.3 Overall Results 67

7.4 Conclusion 74

8 Conclusions and Future Work 75

Bibliography 79

List of Figures

2.1 Schematic illustration of the phenomenon of reverberation. 4

2.2 Comparison between anhechoic and reverberant speech signals, ilustrating the time domain effect of reverberation. 4

2.3 Waveforms and spectrograms of anechoic and reverberant speech signals. 6

2.4 Block diagram of a typical an Automatic Speech Recognition system. 7

2.5 Word error rate of a speech recognition system as a function of the source-microphone distance. The zero distance corresponds to the anechoic speech signals. 8

2.6 Example of room impulse response in the time domain, showing the direct sound followed by early and late reflections. 9

2.7 Diagram of a generic multichannel dereverberation system. 12

3.1 PESQ score versus source-microphone distance for two reverberation times values. 19

4.1 Pareto-optimal set, marked with a solid line, for an scenario with two objectives (O_1, O_2) to be maximized. 22

4.2 Five different solutions shown in the objective space. 23

4.3 Illustrative example of proposed optimization using PESQ and SRMR as quality-assessment measures 25

4.4 Numerical example of the proposed method 26

5.1 Diagram of the algorithm A1. 31

5.2 Q_{MOS}×PESQ relation for the training process of algorithm A1. . . . 34

5.3 SRMR×PESQ relation for the training process of algorithm A1. . . . 34

5.4 NCM×PESQ relation for the training process of algorithm A1. 35

5.5 NCM×PESQ relation for the training process of algorithm A1. 35

5.6 Operating points selected to evaluate the influence of PESQ and Q_{MOS} increase on speech intelligibility. 36

5.7 WER×PESQ relation for different fixed values of Q_{MOS}. 37

5.8 WER×Q_{MOS} relation for different fixed values of PESQ. 38

5.9 Average of WER for each one of the selected points on the Q_{MOS}×PESQ plot. 38
5.10 Some non-dominated solutions located on the Pareto frontier that jointly maximizes the PESQ and Q_{MOS} measures. 39
5.11 Average of WER for candidates to optimal operating point considering the PESQ and Q_{MOS} measures. 40
5.12 Operating points located on the convex hull that jointly maximizes the NCM and PESQ values. 40
5.13 Average of WER for candidates to optimal operating point considering the NCM and PESQ values. 41
5.14 Q_{MOS}×PESQ and SRMR×PESQ plots for the training database, showing the unprocessed, original, and optimized operating points for algorithm A1. 42
5.15 WER results for the whole database comparing the different configurations of algorithm A1. 44
5.16 Q_{MOS}, SRMR and PESQ results for SimData and RealData development dataset comparing the different configurations of algorithm A1. 45
5.17 Q_{MOS}, SRMR and PESQ results for SimData and RealData evaluation dataset comparing the different configurations of algorithm A1. . 46

6.1 Diagram of algorithm A2. 49
6.2 Schematic representation of the RIR division into early and late components. 49
6.3 Q_{MOS}×PESQ and SRMR×PESQ for each group of parameters for the optimization process of algorithm A2. The red circle represents the preferred solution for groups 1 and 2. 55
6.4 Feasible objective spaces of the optimization process of algorithm A2, using Q_{MOS}×PESQ and SRMR×PESQ measures. 56
6.5 Original configurations of algorithm A2 represented in the Q_{MOS}×PESQ and SRMR×PESQ plots. 56
6.6 Candidates for optimal operating points of algorithm A2. 57
6.7 Q_{MOS}×PESQ and SRMR×PESQ plots for the training database, showing the unprocessed, original, and optimized operating points of algorithm A2. 58
6.8 Word error rate (WER) results for the whole database comparing the different configurations of algorithm A2. 60

6.9 Q_{MOS}, SRMR and PESQ results for SimData and RealData development dataset comparing the different configurations of algorithm A2. 61
6.10 Q_{MOS}, SRMR and PESQ results for SimData and RealData development dataset comparing the different configurations of algorithm A2. 62

7.1 Diagram of algorithm A3. 65
7.2 Q_{MOS}×PESQ and SRMR×PESQ for the optimization process of algorithm A3. 67
7.3 Function that maps the Q measure into the Q_{MOS}. 67
7.4 Candidates for optimal operating point of algorithm A3. 68
7.5 SRMR×PESQ plot for the training database, showing the unprocessed, original, and optimized operating points of algorithm A3. . . . 69
7.6 Word error rate (WER) results for the whole database comparing the different configurations of algorithm A3. 71
7.7 SRMR and PESQ results for SimData and RealData development dataset comparing the different configurations of algorithm A3. 72
7.8 SRMR and PESQ results for SimData and RealData development dataset comparing the different configurations of algorithm A3. 73

8.1 WER for the different regions of Q_{MOS} and PESQ. 76
8.2 Comparison of the operating points of algorithms A1, A2 and A3 in the SRMR×PESQ plot. 77
8.3 Example of feasible objective space using the PESQ, SRMR and Q_{MOS} measures. 78

List of Tables

4.1 Quantity of data for development and evaluation sets of SimData and RealData. 28
4.2 Average WER in % obtained for the clean and reverberant speech signals from development and evaluation datasets under every reverberant condition. 28

5.1 Range of values of each parameter used in the enhancement process of algorithm A1. 33
5.2 Results of algorithm A1 for SimData and RealData development datasets. 42
5.3 Results of algorithm A1 for SimData and RealData evaluation datasets. 43
5.4 Comparison of the parameter values and the total average WER between the original and the enhanced configuration of algorithm A1. . 47

6.1 Parameters chosen for the enhancement process of algorithm A2, including a brief description, the equation in which they are involved, their original values, and the range of analysis. 54
6.2 Grouping of the parameters and number of combinations for the enhancement process of algorithm A2. 55
6.3 WER average in % corresponding to the operating points $C1$, $C2$, $C3$, $C4$, and $C5$ using the development dataset. 57
6.4 Comparison between the parameter values for the original and the optimal configurations for algorithm A2. 58
6.5 Results of algorithm A2 for SimData and RealData development datasets. 59
6.6 Results of algorithm A2 for SimData and RealData development datasets. 60
6.7 Comparison of the parameter values and the total average WER between the original and the enhanced configurations of algorithm A2. . 63

7.1 Parameters chosen for the optimization process of algorithm A3, including a brief description, their original values, and the range of analysis. 66

7.2 WER average in % corresponding to operating points $D1$, $D2$, $D3$, $D4$, and $D5$ using the development dataset. 68

7.3 Comparison between the parameter values for the original and the optimal configurations of algorithm A3. 68

7.4 Results for SimData and RealData development datasets of algorithm A3. 69

7.5 Results of algorithm A3 for SimData and RealData evaluation datasets. 70

7.6 Comparison of the parameter values and the total average WER between the original and the enhanced configuration of algorithm A3. . 74

8.1 Comparison of the WER achieved by the original and enhanced versions of the algorithms A1, A2, and A3. 76

List of Symbols

E_d	Direct energy level, p. 10
E_r	Reverberant energy level, p. 10
R_{DT}	Reverberation decay tail, p. 17
T_{60}	Reverberation time, p. 10
σ_r^2	Room spectral variance, p. 10

List of Abbreviations

AIR	Acoustic impulse response, p. 8
ASR	Automatic speech recognition, p. 6
BSD	Bark spectral distortion, p. 18
CD	Cepstrum distance, p. 18
DFT	Discrete Fourier transform, p. 5
DRR	Direct-to-reverberant ratio, p. 10
DSB	Delay-and-sum beamformer, p. 11
EDC	Energy decay curve, p. 10
FFT	Fast Fourier transform, p. 30
HTK	Hidden Markov model toolkit, p. 28
IDFT	Inverse discrete Fourier transform, p. 65
IIR	Infinite impulse response, p. 17
IP	Internet Protocol, p. 13
ITU-T	International Telecommunications Union, p. 14
LLR	Log-likelihood ratio, p. 18
MFCC	Mel frequency cepstral coefficient, p. 28
MMSE-LSA	Minimum mean square error log spectral amplitude, p. 48
MOS	Mean opinion score, p. 14
MSE	Mean squared error, p. 11
NBP	New Brazilian Portuguese, p. 3

NCM	Normalized covariance measure, p. 19
PESQ	Perceptual evaluation of speech quality, p. 17
REVERB	REverberant Voice Enhancement and Recognition Benchmark, p. 27
RIR	Room impulse response, p. 8
RSV	Room spectral variance, p. 10
RT	Reverberation time, p. 10
SIR	Signal-to-interference ratio, p. 50
SNR	Signal-to-noise ratio, p. 19
SNRseg	Segmental signal-to-noise ratio, p. 18
SRMR	Speech-to-reverberation modulation energy ratio, p. 16
SRR	Signal-to-reverberation ratio, p. 17
STFChT	Short-time fan chirp transform, p. 48
STFT	Short-time Fourier transform, p. 5
STI	Speech transmission index, p. 19
VAD	Voice activity detector, p. 51
WER	Word error rate, p. 20
fwsegSNR	Frequency weighted segmental signal-to-noise ratio, p. 18

Chapter 1

Introduction

When a person is speaking in a certain enclosure, the sound waves will be reflected by several surfaces and objects within the enclosure. In this manner, the receiver or receivers will not perceive only the direct speech signal, but also multiple delayed and attenuated signals created by those reflections. This phenomenon is known as reverberation and it is present in everyday life experience of sound.

Depending on the application context, the effects on speech caused by reverberation may be problematic. For example, in hands-free terminals, human-machine communication systems, videoconferencing, or hearing aids, reverberation can degrade speech intelligibility and perceptual quality. In all these examples, the speaker is normally at a considerable distance from the microphone, so the observed signal can be affected by reverberation caused by reflections from walls, floors, ceilings, furniture and other objects. This situation is worsened in noisy environments.

Nowadays, there is a growing demand for hands-free speech input for various telecommunication systems, owing to the increasing use of portable devices and a worldwide expansion of broadband internet access. These factors are also mixed with and pulled by the development of numerous advanced speech applications, such as automatic speech-to-text conversion, speaker identification, source localization, voice-controlled device operation, car interior communication systems and hearing aids. The alteration of the characteristics of the speech signal caused by reverberation can be problematic for all those signal processing applications. Therefore, speech enhancement algorithms are commonly used to improve the quality and intelligibility of reverberant speech signals at the receiving end. These dereverberation techniques are of great importance, and constitute a topic of study with many important research questions yet unanswered.

1.1 Motivation and Work Proposal

Reverberation effects are known to be a major cause of degradation of automatic speech recognition performance and loss of speech intelligibility. To mitigate these effects, many dereverberation methods have been proposed and developed in the last decades. The objective of this work is to present an enhancement procedure for dereverberation algorithms. The main idea is to fine tune certain parameters of a given algorithm in order to jointly optimize several perceptual-assessment quality measures. By doing so, the intelligibility of the processed speech signals can be improved, thus increasing the algorithm performance at no additional computational cost.

Several quality and intelligibility measures will be analyzed and combined, and the proposed methodology will be evaluated with three different dereverberation algorithms. Since no universally accepted set of measures has been fully established for evaluating dereverberation algorithms, the results in this work will be also assessed and compared trough the word error rate of a speech recognition system, which provides a more objective comparison.

1.2 Organization of the Thesis

The remainder of this thesis is organized as follows:

Chapter 2 introduces the fundamentals of reverberation and its effects in communication systems, particularly in automatic speech recognition systems.

Chapter 3 presents subjective and objective measures that are commonly used to assess the quality and intelligibility of speech signals.

Chapter 4 details the method proposed in this work. A procedure based on the multi-optimization of the assessment measures is presented with the aim of improving the performance of dereverberation algorithms.

Chapters 5, 6, and 7 presents three different dereverberation algorithms that are enhanced using the proposed method.

Finally, Chapter 8 presents the conclusions of the work and possible future steps to extend the results achieved here.

Chapter 2

Fundamentals of Reverberation

This chapter provides the fundamentals necessary to understand the effects of reverberation and the need for dereverberation algorithms. Basic definitions and concepts are presented, providing a brief overview of widely used terms in this work. First, the phenomenon of reverberation and its effects in time and frequency domains are detailed. Furthermore, a simple experiment showing the negative effect of reverberation in automatic speech recognition systems is described. Next, the acoustic response of reverberation enclosures (rooms particularly) is described. Finally, the structure of generic dereverberation systems is presented.

2.1 Reverberation and its Effects

When speech signals are obtained in an enclosed space by one or more microphones positioned at a distance from the talker, the observed signal consists of a superposition of many delayed and attenuated copies of the speech signal due to multiple reflections from the surrounding wall and objects, as illustrated in Figure 2.1. This phenomenon is known as reverberation. When the reflected wave is perceived as a distinct repetition of the sound, it is then called an echo, a different phenomenon in which reflections are not perceived as integrated with the original sound as in the case of reverberation.

The direct path is defined as the acoustic propagation path from the talker to the microphone without reflections. The reflected sound waves suffer delays since their propagation paths are longer than the direct-path. Besides, attenuations of these reflected waves occur due to frequency-dependent absorption phenomenon on surfaces such as floors, ceilings, furniture and other objects. Thereby the resulting reverberant signal received at the microphone is composed by the set of the direct-path and the delayed and attenuated multipath copies.

The effects of reverberation on speech are clearly visible in the waveform representation. Figure 2.2 shows an example time waveform of the sentence "A sensibil-

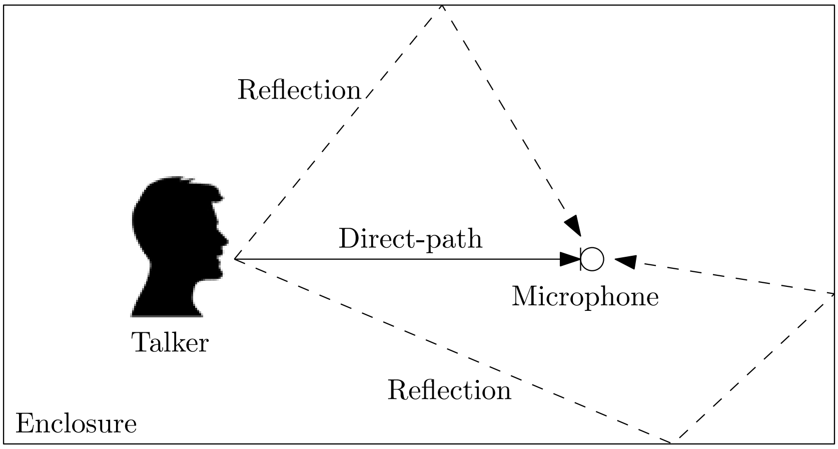


Figure 2.1: Schematic illustration of the phenomenon of reverberation. The reverberant signal recorded at the microphone is composed by the direct-path and the attenuated multipath reflections.

idade indicará a escolha. A Amazônia é reserva ecológica do globo". The speech signal, sampled at 48 kHz, was taken from the new Brazilian-Portuguese (NBP) database [1]. The anechoic (non-reverberant) speech signal is depicted above and a reverberant version is represented below. Due to the smearing of the phonemes in time caused by reverberation, the silence intervals between words and syllables are filled up, and subsequent phonemes overlap [2].

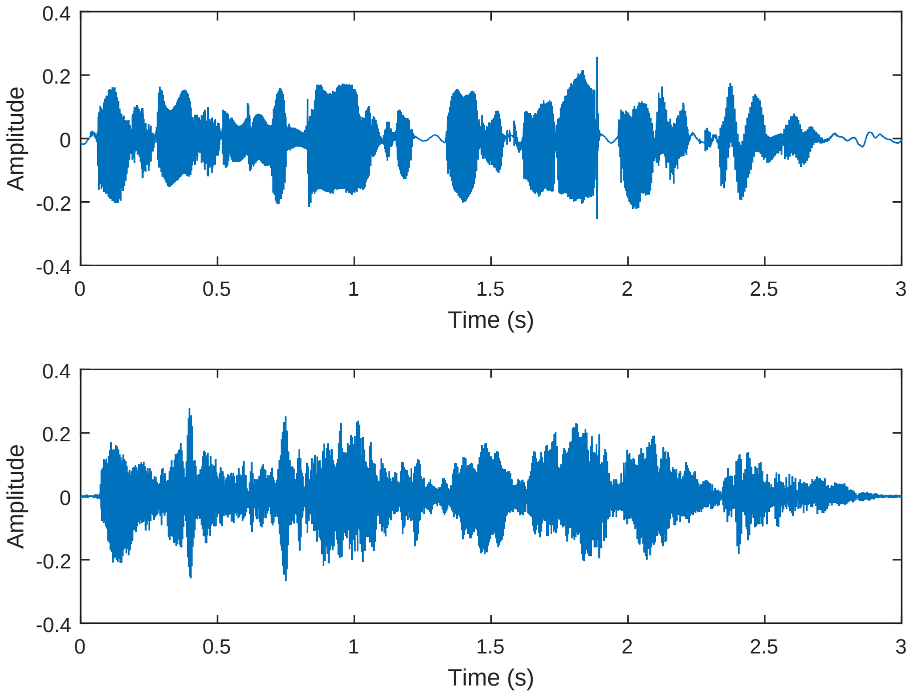


Figure 2.2: Comparison between anhechoic (above) and reverberant (below) speech signals, ilustrating the time domain effect of reverberation.

The reverberant speech signal can be understood as the same source signal coming from several different sources positioned at different locations, therefore arriving with different times and intensities, which adds spaciousness to the sound and gives the perceptual impression of the talker sounding far away from the microphone [3]. These effects, when moderately applied, can add a pleasant sense of the acoustic space to common listeners, but it is almost always unhelpful in voice communication [3]. The alteration of the characteristics of the speech signal caused by reverberation is problematic for signal processing applications such as speech recognition, source localization and hearing aids. The deleterious effects are generally magnified with increasing distance between the talker and the microphone.

2.1.1 Effects of Reverberation in the Frequency Domain

Speech analysis is commonly done in time domain or in frequency domain. Since speech signals have temporal and spectral characteristics changing markedly over time, the discrete Fourier transform (DFT) of an entire speech signal is not appropriate. However, if we consider the analysis of short segments (between 10 and 30 ms) of the speech, the DFT is applicable to speech processing, since the properties of speech do not change much during those segments [4].

A Fourier representation that reflects the time-varying properties of the speech waveform is the short-time Fourier transform (STFT), defined as

$$X(m,\omega) = \sum_{n=-\infty}^{\infty} x(n)w(m-n)e^{-j\omega n}, \tag{2.1}$$

where $x(n)$ is the input signal and $w(n)$ is the analysis window, which is time-reversed and shifted by m samples. In speech analysis the Hamming window is typically used. This window reduces distortions due to its smooth shape [4], and is defined as

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L-1}\right) & 0 \leq n \leq L-1 \\ 0 & \text{otherwise} \end{cases}, \tag{2.2}$$

where L is the window length.

The STFT is a function of two variables: the discrete-time index, m, and the frequency variable, ω. A discrete version of the STFT can be obtained by sampling the frequency variable ω at N uniformly spaced frequencies, that is, at $\omega_k = 2\pi k/N$, for $k = 0, 1, \cdots, N-1$. The resulting discrete STFT is defined as

$$X(m,\omega_k) \triangleq X(m,k) = \sum_{n=-\infty}^{\infty} x(n)w(m-n)e^{-j\frac{2\pi}{N}kn}. \tag{2.3}$$

The spectrogram of a speech signal consists on a two-dimensional display of the power spectrum of speech as a function of time, defined as

$$S(m, \omega) = |X(m, \omega)|^2. \tag{2.4}$$

The spectrogram is a widely used tool for studying the time-varying spectral and temporal characteristics of speech [4]. It describes the speech signal's relative energy concentration in frequency as a function of time, thus reflecting the time-varying and spectral properties of the speech waveform. Spectrograms are typically displayed in gray scale, where dark colors indicate large magnitudes in the spectrum and white colors indicate valleys.

The effects of reverberation on speech are also visible in the spectrogram representation. Figure 2.3 shows the same two waveforms of Figure 2.2, corresponding to a anechoic and a reverberant speech signal, with their respective spectrograms, in which is also noticeable the smearing caused by reverberation.

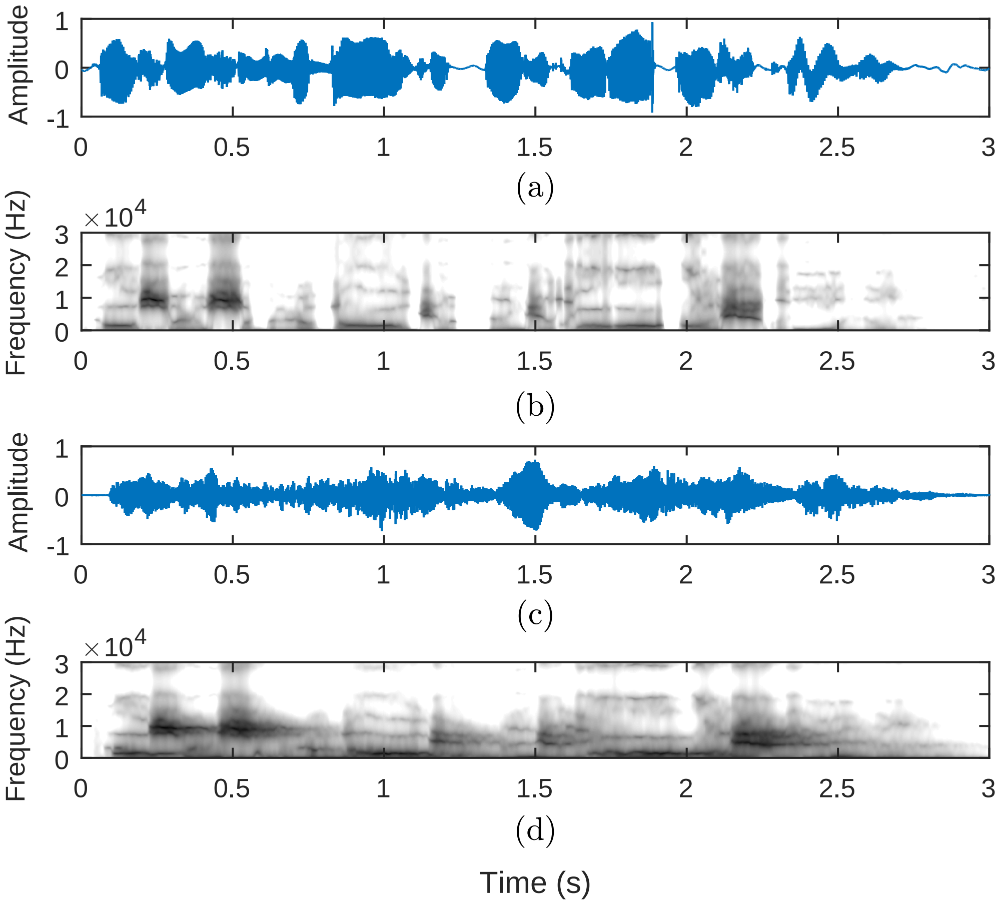


Figure 2.3: Waveforms and spectrograms of (a) and (b) anechoic and (c) and (d) reverberant speech signals, showing the effect of reverberation in the time-frequency domain.

2.1.2 Effects of Reverberation on Automatic Speech Recognition

The performance of automatic speech-recognition (ASR) systems is severely affected by reverberation, and tends to decrease drastically as the source-microphone distance increases [2]. Figure 2.4 shows a block diagram of a typical speech recognition system. First, feature vectors are extracted from the speech signal, as a means to characterize the essential information present in the speech. Next, based on these features, the most likely text is found by the decoder by using two types of knowledge models: an acoustic model, which contains knowledge required to decode the features into phonemes; and the linguistic model, employed to decode these phonemes into text. These models are usually trained before the decoding step. In most cases, the acoustic model is trained on a set of acoustic features extracted from clean speech signals. Thus, the distortion caused by reverberation on the input signal of an ASR system leads to degraded recognition performance.

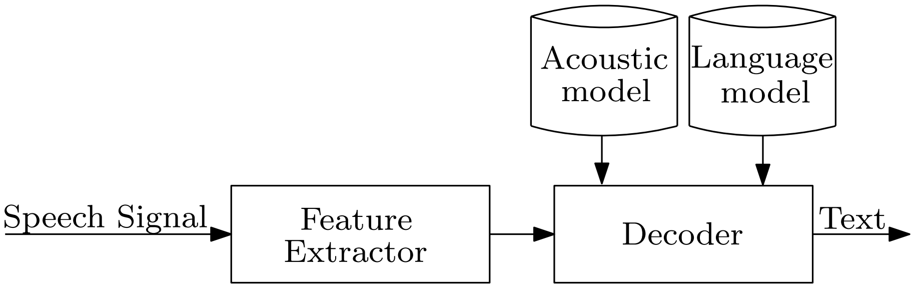


Figure 2.4: Block diagram of a typical an Automatic Speech Recognition system.

A recognition experiment is described here in order to show the influence of reverberation on the performance of a speech recognition system. For this experiment, a set of 247 speech utterances from the WSJCAM0 corpus [5] was employed. The simulated reverberant signals were provided in [6]. For this experiment, three scenarios were devised: the signals without reverberation (i.e., under anechoic conditions); the signals inside a room of length 6.27 m and width 2.59 m with a source-microphone distance of 50 cm (moderate reverberation); and with a source-microphone distance of 200 cm (severe reverberation). The ASR system used here is based on the hidden Markov model toolkit (HTK) [7]. Figure 2.5 shows the average word error rate of the recognition system for the three scenarios, where the clean signals are considered to have zero source-microphone distance. As can be seen, the word error rate increases with increasing source-microphone distance, thus indicating that the effects of reverberation on the ASR system are rather severe.

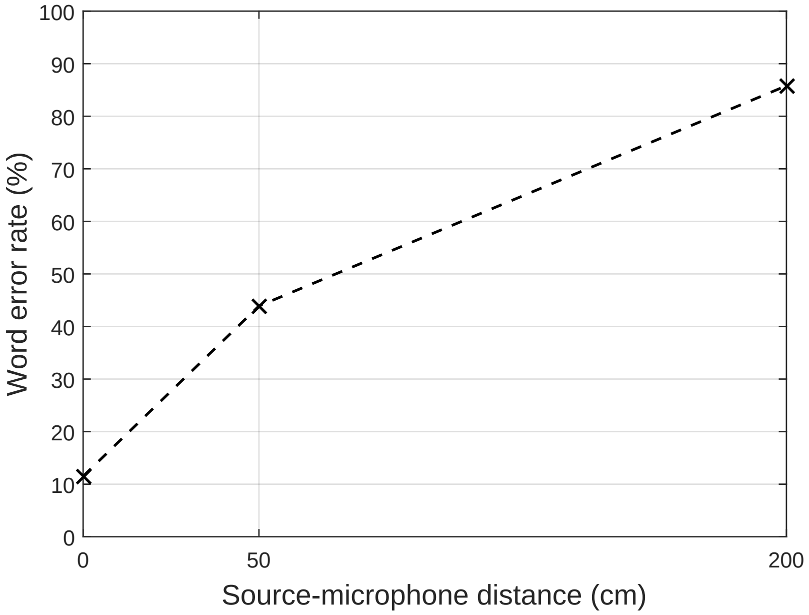


Figure 2.5: Word error rate of a speech recognition system as a function of the source-microphone distance. The zero distance corresponds to the anechoic speech signals.

2.2 Acoustic Impulse Responses

The acoustic impulse response (AIR) characterizes the acoustics of a given enclosure. When the acoustic scenario is limited to be within a room, the impulse response is referred to as a room impulse response (RIR), whose shape depends on factors such as the size of the room, the reflectivity of the surfaces in the room and the talker-to-microphone distance [3]. The enclosure (or room) is considered to be the system that might be responsible for the incidence of reverberation or ambient noise. The input to this system is the speech source signal (e.g. talker or speaker) and the output of this system is the signal received at the microphone. In the absence of reverberation and additive noise, the captured signal is identical to the signal produced by the source, except for the delay due to the propagation from the source to the microphone.

A RIR is generally assumed to consist of three parts: a direct-path response, early reflections and late reverberation. Figure 2.6 shows an example of a simulated RIR, where the initial period of zero amplitude is referred to as the direct-path propagation delay, followed by a peak corresponding to the direct sound. Depending on the source-microphone distance and the reflectivity of the surfaces in the room, the amplitude of this peak varies. In the example illustrated in this figure, the strong direct-path component indicates that the source-microphone distance is relatively short. Early and late reflections are indicated in the figure as two distinct regions of the RIR. Early reflections often corresponds to the initial portion of the impulse response and contains most of its energy. They consist of a set of well defined impulses of large magnitude relative to the smaller magnitude and diffuse nature of late reflections [3]. Early reflections cause spectral changes and lead to a perceptual effect referred to as coloration. Late reflections are referred to as the tail of the impulse response and consist of closely spaced, decaying impulses seemingly randomly distributed. They are known to be a major cause of ASR performance degradation, as well as speech fidelity and intelligibility loss [2].

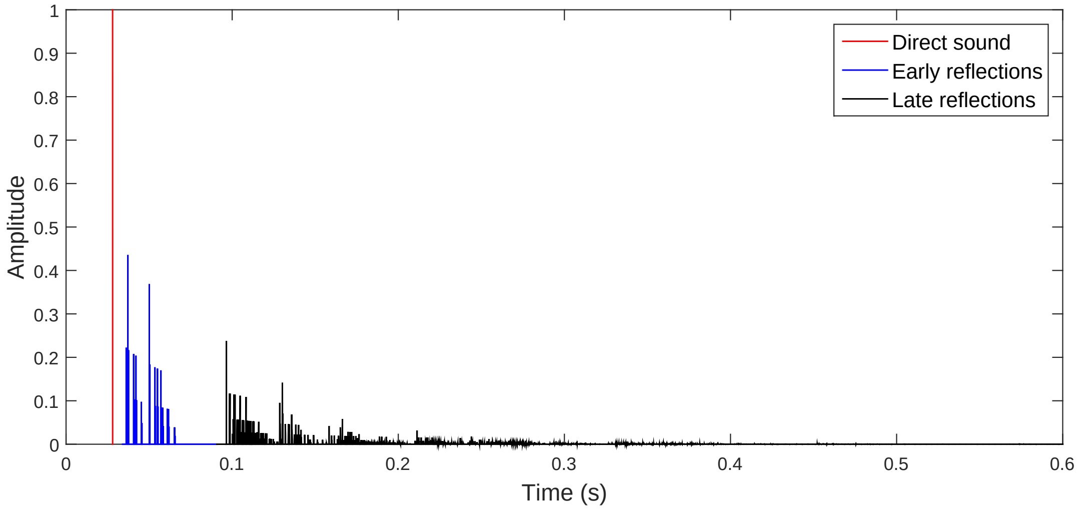


Figure 2.6: Example of room impulse response in the time domain, showing the direct sound followed by early and late reflections.

For a linear, causal and time-invariant RIR, $h(n)$, the reverberant speech signal recorded at the microphone can be expressed as

$$z(n) = \sum_{k=0}^{\infty} h(k)s(n-k), \tag{2.5}$$

where $s(n)$ is the anechoic speech signal. In the discrete STFT domain the previous equation can be expressed as

$$Z(m,k) = H(m,k)S(m,k), \tag{2.6}$$

where $Z(m,k)$, $H(m,k)$ and $S(m,k)$ are the discrete STFTs of $z(n)$, $h(n)$ and $s(n)$, respectively.

In practice, however, the acoustic channel cannot be assumed to be time-invariant, due to changes in source or microphone position, temperature, positioning of room furnishings and movements inside the room [8].

Three often-used measures associated with the reverberation effect and related to the RIR are detailed next. These measures are the reverberation time (RT or

T_{60}), the direct-to-reverberant ratio (DRR), and the room spectral variance (RSV).

2.2.1 Reverberation Time

An often-used quantification of the RIR is the reverberation time, or T_{60}, originally introduced by Sabine [9]. It is defined as the time taken for the reverberant energy to decay 60 dB once the sound source has been abruptly shut off. The reverberation time for a room depends on the room geometry and the reflectivity of all internal surfaces, and it can be considered approximately constant when measured at any location in a given room [3]. In practice, higher T_{60} values indicate more severe reverberation.

Reverberation time can be measured by exciting a room with a broadband signal and recording the resulting decay of the squared sound pressure against time, known as the energy decay curve (EDC), defined as

$$\mathrm{EDC}(t) = \int\limits_{t}^{\infty} h^2(\tau)\mathrm{d}\tau. \tag{2.7}$$

Then, the T_{60} is the required time for the EDC to decrease 60 dB. The reverberation time in typical office-sized rooms can be expected to vary in the range [0.1,1.0] s [3]. Numerous proposals to estimate the reverberation time can be found in the literature [9–15].

2.2.2 Direct-to-Reverberant Ratio

The DRR is defined as the ratio between the direct E_d and reverberant E_r energy levels of the RIR, that is,

$$\mathrm{DRR} = 10\log_{10}\left(\frac{E_d}{E_r}\right) = 10\log_{10}\left(\frac{\sum\limits_{n=0}^{n_d} h^2(n)}{\sum\limits_{n_d+1}^{\infty} h^2(n)}\right) \quad [\mathrm{dB}], \tag{2.8}$$

where n_d is the discrete-time sample index for the direct-path component of the RIR. This measure is inversely proportional to the source-microphone distance and the reverberation time of the room.

In cases where the reverberating system's impulse response is not easy to estimate, blind estimators for DRR measure have been developed. Some of them can be found in [16, 17].

2.2.3 Room Spectral Variance

The RSV is defined as the variance of the energy spectrum of the RIR in dB [18], and characterizes the reverberation effect in the frequency domain. In that sense, if $H(k)$, for $0 \leq k < K$, is the k^{th} discrete Fourier transform (DFT) coefficient of the RIR $h(n)$, the relative acoustic intensity level is defined as

$$I(k) = 10 \log_{10} \left(\frac{|H(k)|^2}{\frac{1}{K} \sum\limits_{k'=0}^{K-1} |H(k')|^2} \right) \quad \text{[dB]}, \tag{2.9}$$

and the RSV is determined by

$$\sigma_r^2 = \frac{1}{K} \sum_{k=0}^{K-1} \left[I(k) - \left(\frac{1}{K} \sum_{k'=0}^{K-1} I(k') \right) \right]^2. \tag{2.10}$$

In [19], one can found a semi-blind algorithm for the RSV estimate, which requires a previous knowledge of the reverberation time and DRR values.

2.3 Dereverberation Systems

For many speech processing applications, such as hearing aid and speech recognition systems, a one-microphone approach for dereverberation is highly desirable. However, the use of microphone arrays is commonly associated to dereverberation techniques found in the literature [20, 21]. A generic system diagram for dereverberation, which can represent both single and multichannel dereverberation, is shown in Figure 2.7. For this generic system, one assumes the presence of additive noise, represented by $\nu_i(n)$, for $i = 1$ to M, where M is the number of channels. Each observed signal, $y_i(n)$, at microphone i, corresponds to the superposition of the direct-path signal and a theoretically infinite set of reflections of the talker speech signal $s(n)$ arriving at the microphone at later time instances with different attenuations and added noise. Thus, it can be considered that the received signal is convolved with a different RIR for each channel.

The objective of dereverberation systems is to output a signal $x(n)$ which is a good estimate of the anechoic signal $s(n)$. There are different criteria for the definition of 'good estimate', from those related to perceptual quality to a direct minimum mean squared error (MSE). Dereverberation systems from the literature can be grouped into three different families:

1. *Speech enhancement*: the received speech signal $y(n)$ is modified in order to improve some features of the anechoic speech signal $s(n)$ according to a de-

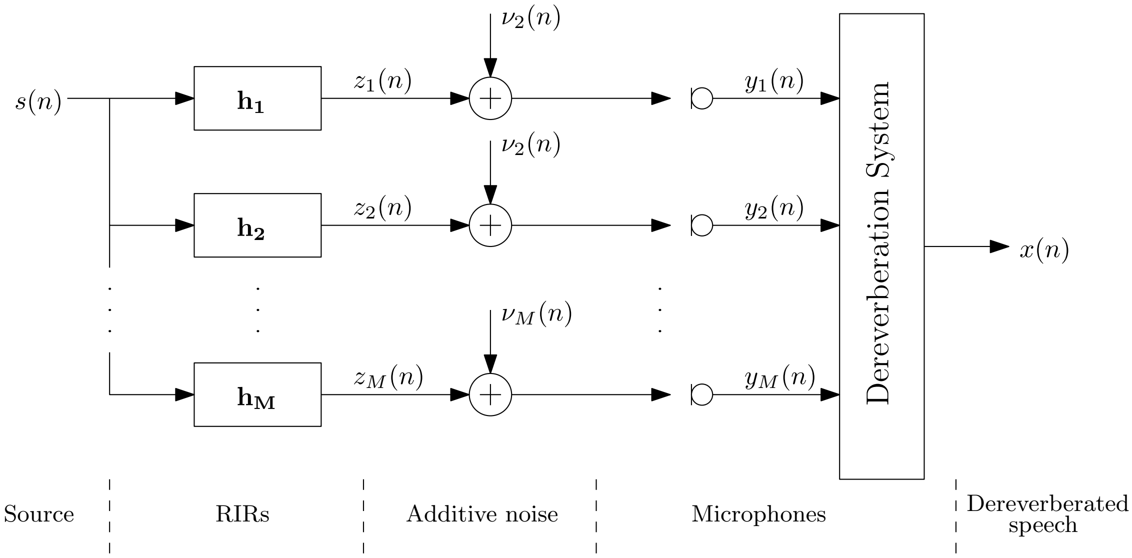


Figure 2.7: Diagram of a generic multichannel dereverberation system.

fined model of the speech waveform or spectrum. Some speech enhancement approaches to dereverberation can be found in [8, 22–31].

2. *Blind deconvolution*: the RIR is identified blindly, that is, using only the observed signals $y_i(n)$. Then, an inverse filter that equalizes the effect of the RIR is designed. This technique often employs multiple microphones in the dereverberation process. Blind acoustic system identification algorithms for dereverberation can be found in [3, 32–38].

3. *Beamforming*: this technique provides the ability for a sensor array to focus on a specific source with a particular angular position with respect to the array. Generally, this method offers better results; however, it is dependent on the availability of multi-microphone inputs, which is not possible in several real communication systems. The most used technique is the delay-and-sum beamformer (DSB), in which the microphone signals are delayed to compensate for different times of arrival, and then weighted and summed. In this way, the components due to direct-path are added while the components due to reverberation are attenuated. Some beamforming techniques for dereverberation can be found in [20, 21, 39–41].

In the next chapter, the evaluation of the performance of dereverberation algorithms will be discussed.

Chapter 3

Evaluating the Performance of Speech Dereverberation Algorithms

Many speech enhancement algorithms have been developed to enhance the quality and intelligibility of distorted speech. Their performance can be assessed using either subjective listening tests or objective measures. By comparing the speech quality before and after processing by the algorithm being evaluated, it is possible to investigate the speech quality improvement.

Objective measures can be classified into two groups: reference and no-reference. The reference measures compare the distorted signal with the undistorted signal (i.e., the reference signal), which needs to be available. The no-reference measures do not use a reference signal, i.e., the speech quality is determined given only the distorted speech signal. For the objective measures to be valid, they need to correlate well with subjective listening tests. Some of the existing objective measures have been adopted to characterize the perceived effect of reverberation in a speech signal.

In this chapter, subjective and objective measures that can be used to determine the enhancement of speech quality will be discussed and analyzed, particularly useful to determine the dereverberation effect. Some knowledge regarding those measures is necessary to understand the method proposed in this work, which is presented in the next chapter.

3.1 Quality and Intelligibility of Speech

Speech quality and speech intelligibility are not synonymous terms. Hence, different assessment methods are used to evaluate the quality and intelligibility of processed speech. Quality is highly subjective in nature and it is difficult to evaluate reliably.

On the other hand, intelligibility can be easily measured and quantified by counting the number of words or phonemes that are correctly identified, either by real listeners or by ASR systems. The relationship between speech quality and speech intelligibility is not fully understood, and the acoustic link between them has not been yet identified [4]. Speech can be highly intelligible, yet of poor quality, and vice-versa. For example, a sine-wave speech [42][1] has bad quality since it is perceived as being 'tonal' and mechanical sounding, yet it can be highly intelligible. Conversely, speech can also have good quality and yet not be completely intelligible. For example, speech signals transmitted over IP networks with severe packet loss may become barely intelligible since certain words may be missing, even if the perceived quality of the remaining words is quite high. Even so, in most cases, it is expected that an improvement in the quality of a speech signal implies an improvement in its intelligibility.

Depending on the application at hand, it may be more beneficial to aim at the increase in intelligibility rather than in quality, or conversely. For example, in automatic speech-to-text conversion applications, it is desirable to attain high intelligibility of the processed speech, regardless of the speech quality. In hearing-aid applications, however, it is desirable that the speech enhancement algorithm preserves or enhances not only speech intelligibility but also speech quality. When the degradation of speech is caused by reverberation, it is important to note that speech fidelity and intelligibility are mostly degraded by late reverberation. On the other hand, the coloration effect caused by the early reflections reinforces the direct sound, and is therefore considered useful regarding speech intelligibility [3].

3.2 Subjective Measures

Subjective speech quality and intelligibility measures can be obtained using listening tests with human participants. In the case of speech intelligibility, such tests generally fall into three main classes: recognition of nonsense syllables, recognition of single meaningful words, and recognition of meaningful sentences. In most of them, speech intelligibility is quantified in terms of percentage of words (or syllables) identified correctly [4]. For the evaluation of speech quality, listeners are presented with recordings that are enhanced by a certain algorithm, and asked to rate the quality of each signal on a numerical scale, typically a 5-point scale with one indicating poor quality and five indicating excellent quality. The measured quality of the test signal is obtained by averaging the scores obtained from all listeners. This average score is

[1] Sine-wave speech is an intelligible synthetic acoustic signal composed of three or four time-varying sinusoids. Together, these few sinusoids replicate the estimated frequency and amplitude pattern of the resonance peaks of a natural utterance.

commonly referred to as the mean opinion score (MOS), and is one of the methods recommended by the International Telecommunications Union (ITU-T) [43].

To obtain a realistic variability in the listening test, a large number of subjects is required. Although subjective-listening tests provide perhaps the most reliable method for assessing speech quality or speech intelligibility, these tests can be time-consuming and are unsuitable for real-time applications. Hence, it is highly desirable to devise objective measures of speech quality and intelligibility.

3.3 Objective Measures

Objective quality and intelligibility measures are valuable assessment tools during the design and validation stages of speech enhancement algorithms, codecs, and communication systems. Most objective measures are based on psychoacoustics considerations and trained on subjective databases to represent human perception. A large number of objective measures has been devised for different types of applications. Next, a subset of measures that have been found to correlate well with reverberant environments is described.

3.3.1 Reverberation-Oriented Quality Measures

Some objetive measures have been developed to determine the amount of reverberation present in speech signals. Reliable quantitative measurement of the level of reverberation in a speech signal is particulary difficult and no universally accepted set of instrumental measures has yet been fully established for evaluating dereverberation algorithms. Many of the commonly used measures, the so-called channel-based measures, need information about the channel impulse response. When this information is not available, estimates must be used in order to compute the measures. For example, in the case of reference measures (that is, when the anechoic speech signal is available), a deconvolution process between the anechoic and reveberant speech signal can be made in order to estimate the RIR. If a direct RIR estimation cannot be computed, blind approaches for the parameters required by the measures must be employed. A set of measures used for the quantitative characterization of reverberation are summarized below.

3.3.1.1 Allen's Measure

This measure estimates the *subjective preference* of reverberant speech and combines the reverberation time (T_{60}) and the RSV (σ_r^2) as follows [44]:

$$P = P_{max} - \sigma_r^2 T_{60}, \tag{3.1}$$

where P is the subjective preference in some arbitrary units and P_{max} is the maximum possible preference. According to this formula, decreasing either the RSV or the reverberation time results in an increased subjective preference (quality) of the speech.

3.3.1.2 QAreverb

The QAreverb measure Q proposed in [1] incorporates the DRR to the Allen's measure as given by

$$Q = -\frac{T_{60}\sigma_r^2}{\mathrm{DRR}^{\gamma}}, \tag{3.2}$$

where the exponent γ sets the importance of the DRR with respect to the other two parameters, and its value is determined empirically in a training stage. In the next stage, the value of Q is mapped onto the MOS scale, yielding the Q_{MOS} measure, using a third-order polynomial of the form

$$Q_{MOS} = x_1 Q^3 + x_2 Q^2 + x_3 Q + x_4, \tag{3.3}$$

where the coefficients x_1, x_2, x_3, and x_4 are determined during the system training. This procedure is followed by a linear-scale adjustment of Q_{MOS} to the grade scale of a distinct subjective test, in order to reduce the MSE between subjective and objective scores:

$$Q_{MOS} = \alpha Q_{MOS} + \beta, \tag{3.4}$$

with α and β possibly determined from some data subset.

A no-reference version of the QAreverb measure can be obtained using blind estimators for T_{60} [14], DRR [17], and RSV [19].

3.3.1.3 Speech to Reverberation Modulation Energy Ratio

The speech-to-reverberation modulation energy ratio (SRMR) [16] is a no-reference signal-based measure defined as

$$\mathrm{SRMR} = \frac{\sum\limits_{k=1}^{4} \bar{\varepsilon}_k}{\sum\limits_{k=5}^{K^*} \bar{\varepsilon}_k}, \tag{3.5}$$

where $\bar{\varepsilon}_k$ is the average energy in the modulation frequency band index k, and the upper summation bound K^* in the denominator is adapted to the speech signal under test. This measure is computed by performing spectral analysis on the modulation envelopes of the speech signal and is calculated as the ratio of the average energy in the low modulation frequencies (4 – 18 Hz) (attributed mostly to spoken speech components) to the high modulation frequencies (29 – 128 Hz) (which are mostly attributed to noise and room acoustic effects). Larger values of SRMR are assumed to indicate better speech quality.

3.3.1.4 Other Reverberation-Oriented Measures

Many other quality measures that concentrate on measuring the dereverberation effect can be found in the literature. To cite some examples, the signal-to-reverberation ratio (SRR) [45] and the reverberation decay tail (R_{DT}) [46] are widely used measures of reverberant and dereverberated speech.

3.3.2 General Purpose Quality Measures

Most of the objective quality measures of speech quality were developed for the purpose of evaluating the distortions introduced by speech codecs and/or communication channels. These overall quality measures were not designed to determine the quality of reverberant speech, but other distortions that are perceptually important. Next, a subset of general purpose quality measures that are also suitable for evaluating the quality of speech enhanced by dereverberation algorithms are introduced. Among these measures, the ITU-T has standardized the perceptual evaluation of speech quality (PESQ) [47], which is described in detail.

3.3.2.1 Perceptual Evaluation of Speech Quality

PESQ was conceived to predict the listening quality of a speech signal (sampled at 8 KHz) degraded by codecs, background noise and packet loss. This reference measure compares an original (clean) signal with a degraded or enhanced version, and the output is a prediction of the perceived quality it would have been attributed in a subjective listening test. The main details involved in PESQ computation are given below.

1. *Pre-processing*: The clean and degraded signals are first level equalized to a standard listening level, and filtered by a filter with response similar to a standard telephone handset.

2. *Time alignment*: The signals are aligned in time to correct for time delays.

3. *Auditory transform*: In order to account for the distortions that are actually perceived by human listeners, the model transforms the two aligned and filtered signals from the time-amplitude domain into a frequency-loudness domain.

4. *Disturbance processing*: By subtracting the two signal representations an estimate of the audible differences is derived. The audible differences are accumulated over time while they are weighted differently depending on whether a distortion was added to the signal or if parts of the signal were missing after the transmission.

5. *Linear-mapping*: The final PESQ score is computed as a linear combination of the disturbances, providing scores in the range -0.5 to 4.5.

A recommendation for wideband extension to PESQ documented in [48], for speech signals sampled at 16 kHz, included two small changes to the PESQ implementation. First, the filter used originally for modeling the response of telephone headsets was removed. Instead, an infinite impulse response (IIR) filter with a flat response above 100 Hz is used. Second, the PESQ raw output values are mapped using a logistic-type function to better fit the subjective MOS scores. Another extension to PESQ, known as EW-PESQ, was devised in [49] for speech signals sampled at 48 kHz.

An experiment was devised in order to study the relation between PESQ and some basic parameters of reverberation, namely the reverberation time and the source-microphone distance. The speech signals used in the experiment were taken from the NBP database, and were downsampled from 48 kHz to 16 kHz. These signals were obtained by playing and recording anechoic signals in rooms with different reverberating characteristics. For this experiment, the wideband PESQ was used. Figure 3.1 shows the relation between the PESQ score and the source-microphone distance for two different reverberation times. It can be seen that the PESQ score is inversely proportional to the reverberation time and decreases for long source-microphone distances.

3.3.2.2 Other General Purpose Quality Measures

There exist other measures that have been found to be moderately well suited to the assessment of dereverberation algorithms. Some examples of such measures are the Bark spectral distortion (BSD) [50], the segmental signal-to-noise ratio (SNRseg) [51], the frequency-weighted segmental SNR (fwsegSNR) [52], the cepstrum distance (CD) [53] and the log-likelihood ratio (LLR) [4]. For when the clean signal is not available, several no-reference measures have also been proposed in the literature [54–59].

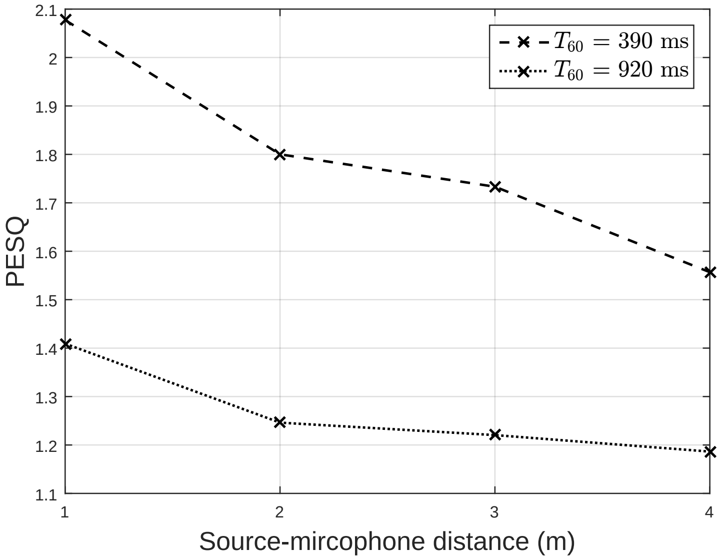


Figure 3.1: PESQ score versus source-microphone distance for two reverberation times values.

3.3.3 Speech Intelligibility Measures

Most intelligibility measures are based on the assumption that intelligibility depends on the audibility of the signal in each frequency band [4]. Audibility is often expressed in terms of SNR, and bands with positive SNR contribute to intelligibility. Thus, objective speech intelligibility scores are predicted based on linear combinations of band SNRs appropriately weighted by some functions. The computation of each band SNR differs across the measure proposed and depends on the background (reverberation, additive noise, etc.) and type of processing. Depending on the method used to compute the SNR, different intelligibility measures have been developed. A powerful and widely accepted family of measures that predict the effect of room acoustics on speech intelligibility are the speech transmission index (STI) [60] measures, which quantify the speech intelligibility in terms of the spectral content of the signal envelope. One of these speech-based STI measures is the normalized-covariance measure (NCM), which has been shown to correlate highly with the intelligibility of reverberant speech [61].

3.3.3.1 Normalized-Covariance Measure

The NCM [62] uses the Pearson's correlation coefficient r between the input and output envelope signals to compute the SNR in each band, using the following equation:

$$\mathrm{SNR}_i = 10\log_{10}\left(\frac{r_i}{1-r_i^2}\right), \tag{3.6}$$

where the subindex i represents the ith band. Following the SNR computation and limitation to the interval [-15,15], the STI is computed by linearly mapping the SNR_i values between 0 and 1,

$$\text{STI}_i = \frac{\text{SNR}_i + 15}{30}. \tag{3.7}$$

Finally, the NCM score is computed as

$$\text{NCM} = \frac{1}{\sum_{i=1}^{K} W_i} \sum_{i=1}^{K} W_i \cdot \text{STI}_i, \tag{3.8}$$

where W_i are the band-importance weights applied to each of the K bands (K is often equal to 20).

3.3.3.2 Word Error Rate

The word error rate (WER) is a commonly used metric to evaluate the performance of ASR systems. It is a measure of the average number of word errors taking into account three error types: substitution (the reference word is replaced by another word), insertion (a word is hypothesized that was not in the reference), and deletion (a word in the reference transcription is missed). Thus, the WER is defined as

$$\text{WER} = \frac{N_S + N_D + N_I}{N}, \tag{3.9}$$

where N_S is the number of substitutions, N_D is the number of deletions, N_I is the number of insertions, and N is the number of words in the reference. Given this definition, the percent word error can be more than 100%. In order to compute this measure, the transcription of all words must be available.

3.4 Conclusion

In this chapter some frequently used objective speech quality and intelligibility measures that are useful to determine the dereverberation quality were analyzed and discussed. Besides, some measures that were not explicitly developed to determine the speech quality in reverberant environments but are sensitive to other important distortions were also introduced. Among these measures it was demonstrated that PESQ is also suitable for the assessment of reverberant signals. The proper combination of the analyzed measures is the key factor of the proposed enhancement method for dereverberation algorithms that this work proposes, as will be detailed in the next chapter.

Chapter 4

Proposed Methodology for Enhancing Dereverberation Algorithms

So far, it has been shown how reverberation distorts speech signals, degrades their quality and adversely affects the performance of ASR systems. To compensate for such detrimental effects, algorithms that mitigate reverberation have become necessary for a large number of applications. However, the process of dereverberation is a difficult and often ill-conditioned problem, and may introduce objectionable artifacts to the processed speech signals. The use of subjective and/or objective quality and intelligibility measurement methods is necessary in order to evaluate the performance of dereverberation algorithms.

In the previous chapter, numerous assessment measures of quality and intelligibility of speech signals were analyzed. This chapter proposes an enhancement strategy for dereverberation algorithms, based on the simultaneous optimization of several of those measures. Basic notions of multi-objective optimization are first introduced. Besides, in order to carry out the work proposal, some practical considerations are also included in this chapter, such as the reverberant-speech database and the speech recognition system used in this work, and setups of the measures used in practice.

4.1 Multi-Objective Optimization and Pareto-Optimal Solutions

Optimization is a procedure of finding and comparing feasible solutions until no better solution can be found. When an optimization problem involves more than one objective, the task of finding one or more optimum solutions is known as multi-objective optimization. In a trivial case, there exists a solution that simultaneously optimizes all objectives. However, such a solution is hardly found. Different solutions may produce trade-offs (conflicting scenarios) among different objectives. Thus, in problems with more than one conflicting objective, there is no single optimum solution, but a number of solutions which are all equally optimal.

Pareto optimality is a concept in multi-objective optimization that allows for the optimization of multiple criteria, enabling all trade-offs among optimal combinations of multiple criteria to be evaluated. Figure 4.1 illustrates a simple case of maximizing two objectives simultaneously (O_1, O_2), with the solid line indicating the Pareto optimal frontier, whereby any improvement with respect to O_1 comes at the expense of O_2. Each point along that frontier represents a unique model parameterization and/or model structure, so Pareto optimality identifies multiple Pareto optimal solutions. Through this procedure one is able to investigate differences among the multiple optimal solutions that are able to optimize varying combinations of assessment criteria. It is worth pointing out that there exist multiple Pareto-optimal solutions in a problem only if the objectives are conflicting to each other. If the objectives are not conflicting to each other, the cardinality of the Pareto-optimal set is one, that is, the optimal solution corresponding to any objective is the same. In the presence of multiple Pareto-optimal solutions, it may be difficult to prefer one solution over the other without any further information about the problem. If higher-level information is available, this can be used to make a biased search. Thus, in a multi-objective optimization, the effort must be made in finding the set of Pareto-optimal solutions.

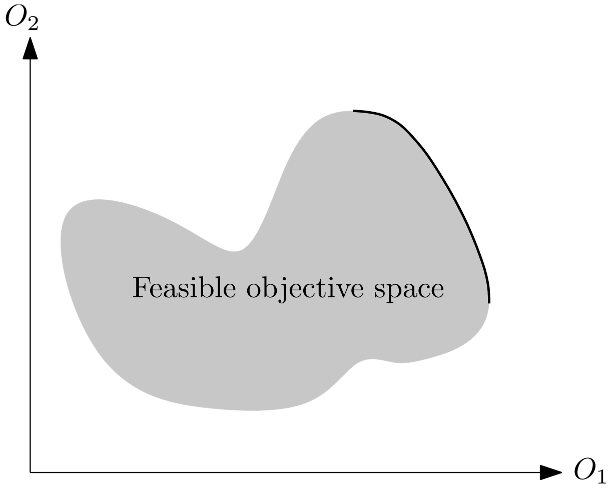


Figure 4.1: Pareto-optimal set, marked with a solid line, for an scenario with two objectives (O_1, O_2) to be maximized.

Most multi-objective optimization methods use the concept of domination in their search. One solution A is said to dominate other solution B if the next two conditions are true:

1. The solution A is no worse than B in all objectives.

2. The solution A is strictly better than B in at least one objective.

If any of the above conditions is violated, the solution A does not dominate the solution B. Let consider the example ilustrated in Figure 4.2, in which the objective O_1 needs to be maximized while the objective O_2 needs to be minimized. Five solutions $\{A, B, C, D, E\}$ are shown in this figure. Using the definition of domination is it possible to decide which solution is better among any two given solutions in terms of both objectives. For example, comparing solutions A and B it can be observed that solution A is better than solution B in both objectives. Thus, the conditions of dominance are satisfied and it can be said that solution A dominates solution B. If solutions A and E are compared, for instance, it can be seen that solution E dominates solution A, since solution E is better in the first objective and no worse than solution A in the second objective. If solutions C and E are now compared, it can be seen that solution E is better than solution C in the first objective, but worse in the second objective. Thus, the first condition of dominance is not satisfied and it cannot be concluded either that solution E dominates solution C, or vice versa, so both solutions are non-dominated.

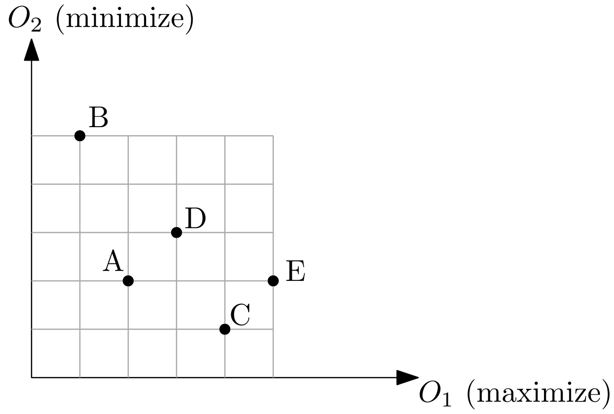


Figure 4.2: Five different solutions shown in the objective space.

Although a two-objective problem is illustrated above, the concept of dominance can be applied in problems with more than two objectives. For a given finite set of solutions, all possible pair-wise comparisons can be performed to search if a solution dominates another and the ones that are non-dominated with respect to each other. At the end, there can be found the so-called *non-dominated* set for the given set of solutions. In the example above, solutions C and E constitutes the non-dominated set of the given set of five solutions. This non-dominated set is precisely the Pareto-optimal set. Different procedures for finding a non-dominated set can be found in [63].

4.2 Proposed Method

The design of dereverberation algorithms often requires the tuning of certain parameters, which are usually chosen so that the final result optimizes one or several measures. The choice of these measures is not a straightforward task. Some reverberation-oriented measures can give a global and quantitative indication about the presence of reverberation, but may not reveal any information about speech quality and intelligibility. Therefore, it is advisable to combine them with other quality-oriented measures.

The proposed method is intended to optimize simultaneously several complementary assessment measures to improve the performance of dereverberation algorithms, using the concepts of multi-objective optimization introduced in the previous section. The core idea is to combine the ability to quantify the reverberation effect inherent to reverberation-oriented measures, such as QAreverb or SRMR, with the ability of general purpose measures, like PESQ, to evaluate the overall quality of speech signals in the presence of other distortions. This combination of measures leads to a multi-objective optimization problem, which can be solved by searching for the Pareto-optimal set in a multi-objective (measure) space. Among the Pareto-optimal solutions, the preferred solution (that is, the new operating point for the dereverberation algorithm) may depend on the application in hand. If, for example, the dereverberation algorithm is intended to improve speech intelligibility, the preferred solution could correspond to that with lower WER, using an ASR system to compare the WER for different optimal (non-dominated) solutions. A six-step procedure for the method proposed in this work is described next:

1. Choose a small but representative training set of reverberant speech signals.

2. Choose a set of relevant parameters of the dereverberation algorithm.

3. Combine the parameters values within a certain range and, for every combination, process the signals from the training set.

4. For each configuration of the algorithm (set of parameter values), compute the selected measures (objectives which are to be maximized or minimized) of the processed signals and average them for the whole training set. This will produce a feasible objective space of points representing the measures values for every setup.

5. Find the Pareto-optimal set (non-dominated solutions) within the feasible objective space.

6. Choose a new operating point for the algorithm from the Pareto-optimal set, using the WER of an ASR system as higher-level decision.

Figure 4.3 shows schematically the proposed procedure, considering a two-objective maximization problem.

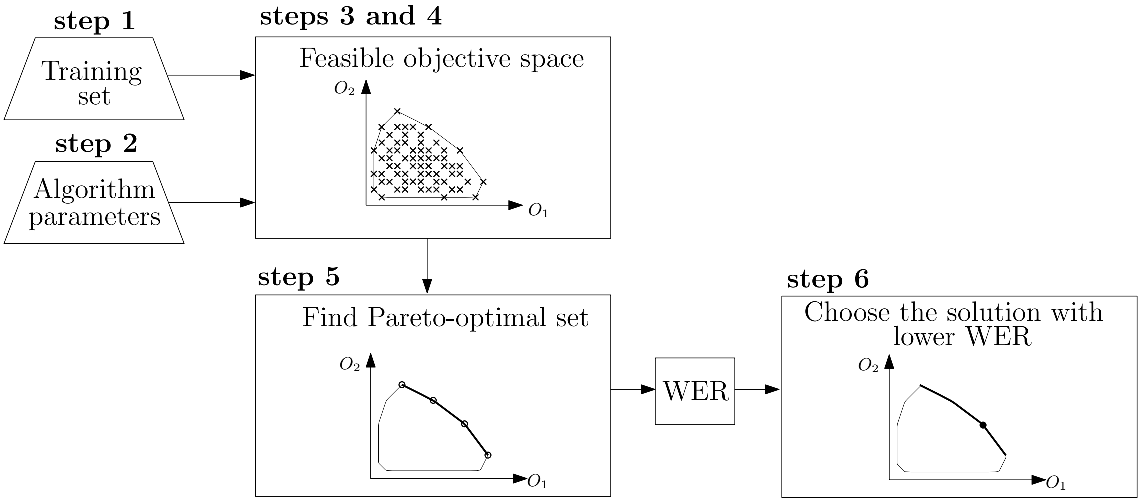


Figure 4.3: Illustrative example of proposed optimization using PESQ and SRMR as quality-assessment measures.

The proposed method can be computationally expensive if the number of parameters is high or the training set is large. Let for example assume that the training set is composed of 15 signals, and 6 parameters of the dereverberation algorithm are chosen, all with an analysis range of 10 values. To carry out this proposal, it would be necessary to process and obtain the measures for $10^6 \times 15 = 15000000$ signals, which is highly time-consuming and perhaps unfeasible. Thus, it is important to limit the size of the training set, the number of parameters and their range of analysis. Another option is to group the parameters into smaller groups and apply the procedure for each one. If in the earlier example the parameters are equally divided into two groups, it would be necessary to process $2 \times 10^3 \times 15 = 30000$ signals, which is more feasible. In compensation, this approach is likely to be less effective in terms of optimization, since possible correlations between parameters are disregarded.

Due to computational limitations, the proposed method might not lead to a global optimization, becoming a sub-optimization problem. Thus, to approximate the optimal solution, it is important to skilfully select the algorithm parameters, knowing their possible correlations, and combine their values within a controlled and appropriate range.

The choice of quality measures has also a significant role. If, for example, two highly correlated measures are selected, their joint optimization would be equivalent to their individual optimizations, which would entail an unnecessary waste of resources.

Now we consider a simple and unreal numerical example of the proposed methodology. The dereverberation algorithm for this example has two parameters: a and b, whose original values are $a = 1$ and $b = 1$. Following the proposed method, these parameters are combined within the range $\{1, 2, 3\}$, which leads to 9 different algorithm configurations. For each of them, the algorithm processes the signals from a certain training set, computing and averaging the SRMR and PESQ measures, which are to be maximized. Figure 4.4 shows their hypothetical objective feasible set, in which every algorithm configuration $\{a, b\}$ is marked on the scattered circles indicating the different solutions. After finding the Pareto-optimal set that lies in the Pareto frontier (represented by a dashed line), it can be observed that the original configuration of the algorithm $\{1, 1\}$ is a dominated solution, so it can be optimized. The next step would be to process a set of speech signals for the three optimal configurations and to compute the WER using a certain ASR system. Suppose that the solution $\{1, 2\}$ gives a WER of 40%, the solution $\{2, 3\}$ gives a WER of 20% and the WER of solution $\{2, 2\}$ is 30%. Thus, with the proposed method, the new operating point of the algorithm would correspond to the solution with lower WER ($a = 2, b = 3$). This new configuration is expected to improve the algorithm performance in terms of speech intelligibility.

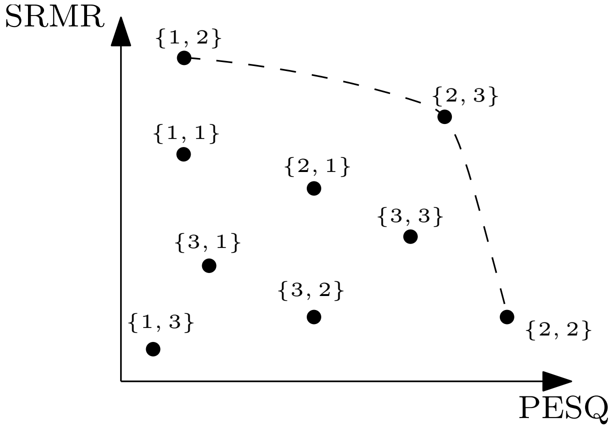


Figure 4.4: Numerical example of the proposed method. Each circle corresponds to one combination of the parameters $\{a, b\}$ of a certain dereverbearation algorithm. The set of all points forms the feasible objective space, where the objectives to be maximized are the PESQ and the SRMR.

4.3 Practical Considerations

The optimization strategy proposed above may be applied to any dereverberation algorithm, with any number of sensors, and with distinct assessment measures. Its effectiveness, however, is illustrated in this work based on three one-microphone dereverberation algorithms described in the following chapters and on QAreverb, SRMR, PESQ, and NCM, discussed in Chapter 3. The setup of the employed measures is described in this section. Also, all experimental data employed in the process is detailed next.

4.3.1 Reverberant Speech Databases

The main database used in this work was provided by the REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge 2014 [6, 64], which divided the data into the so-called development database and evaluation database. Each of these databases were further divided into two datasets:

- SimData: contains speech signals from the WSJCAM0 database [5], artificially convolved with RIRs measured in three different rooms with different volumes (small, medium and large) and two different source-microphone distances (near = 50 cm and far = 200 cm). Background noise was added to each signal at fixed SNR of 20 dB. The reverberation times for the small (Room 1), medium (Room 2) and large (Room 3) rooms are $\{250, 680, 730\}$ ms, respectively. The anechoic signals from this dataset are available.

- RealData: contains a set of real recordings from the MC-WSJ-AV database [65] made in a reverberant and noisy meeting room (Room 4) with two different source-microphone distances (near $\approx$ 100 cm and far $\approx$ 250 cm). The reverberation time for this room is about 700 ms. Since this dataset is composed of real recordings, the respective anechoic signals are not available.

This setup of real recordings (RealData) and simulated data (SimData) allows to evaluate the dereverberation algorithms in terms of both practicality and robustness in a broad range of reverberation conditions. All utterances considered were captured with single-channel microphones at a sampling frequency of 16 kHz. The total size and number of speakers for the SimData and RealData datasets are summarized in Table 4.1. For SimData, the original development and evaluation sets of the WSJCAM0 are divided into three subsets, one for each room. The near and far conditions for each room are based on the same allocated subset. Thus, the size of SimData is twice that of the original WSJCAM0 dataset. For RealData, the development and evaluation sets of the original MC-WSJ-AV are divided into near and far conditions. The total size of the entire database is 4211 reverberant signals.

4.3.1.1 Training Set

A small training set was elaborated in order to substantiate the proposal of this work. This set is composed by one female and one male utterances, randomly selected

Table 4.1: Quantity of data for development and evaluation sets of SimData and RealData.

	SimData		RealData	
	Development	Evaluation	Development	Evaluation
Number of sentences	1484	2176	179	372
Number of speakers	10	28	5	10

from each reverberation condition from the SimData development dataset, totaling 12 reverberant signals. Since most of the used measures are intrusive (i.e., they need access to the anechoic signals), the training set does not contains utterances from the RealData dataset.

4.3.2 Speech Recognition System

In order to have a common basis for evaluating and comparing different approaches of dereverberation algorithms, a baseline speech recognition system [7] was used, which is based on the hidden Markov model toolkit (HTK) [66]. This system follows the structure depicted in Figure 2.4, using Mel-frequency cepstral coefficients (MFCCs) [67] as features. The acoustic model considered here is trained using the clean signals. Table 4.2 summarizes the averaged WER obtained for the clean and reverberant signals under every condition using the speech recognition system for both development and evaluation datasets. As expected, the WER gets worse when reverberation conditions are more severe (i.e., with the increase of reverberation time and source-microphone distance).

Table 4.2: Average WER in % obtained for the clean and reverberant speech signals from development and evaluation datasets under every reverberant condition.

		SimData						RealData	
		Room 1		Room 2		Room 3		Room 4	
Clean	Development	10.50		11.51		10.81		–	
	Evaluation	12.84		12.49		12.13		–	
		Near	Far	Near	Far	Near	Far	Near	Far
Reverberant	Development	15.29	25.29	43.90	85.80	51.95	88.90	88.71	88.31
	Evaluation	18.06	25.38	42.98	88.2	53.54	88.04	89.72	87.34

4.3.3 Objective Assessment Measures

The measures used for the implementation of the proposed method and their configurations are detailed next. All these measures were tested using MATLAB®.

- **QAreverb**: To compute the Q measure of Equation 3.2, the value of the constant factor γ was set to $\gamma = 0.3$, according with [1]. The parameters σ_r^2, T_{60} and DRR are obtained directly from the RIR, which is estimated from the deconvolution process between the clean and the reverberant speech signals. For the case of the RealData dataset, in which the clean signals are not available, a blind version of QAreverb was used, using blind estimators of the parameters as detailed in [17]. The coefficients used for mapping the Q measure into the Q_{MOS} from Equation 3.3 were $x_1 = 0.0014, x_2 = 0.0570, x_3 = 0.6985$, and $x_4 = 4.5390$, and the final linear adjustment of Equation 3.4 was made using $\alpha = 1$ and $\beta = -1.25 \times 10^{-9}$. The motivation behind the choice of these values can be found in [1].

- **SRMR**: The SRMR scores are calculated as per [16].
- **PESQ**: The PESQ scores are calculated using the wideband implementation of the PESQ measure provided in [4].
- **NCM**: For the NCM, the following weighting function was used in Equation 3.8:

$$W_j = \left(\sum_n x_j^2(n) \right)^p, \tag{4.1}$$

where $x_j(n)$ denotes de envelope of the target signal in the jth band and p was set to 1.5. The implementation of the NCM is provided in [4].

4.4 Conclusion

In this chapter, the proposed methodology to enhance dereverberation algorithms was introduced. The enhancement strategy consists on tuning certain parameters of a given algorithm so that several objective measures are optimized simultaneously. After finding the non-dominated solutions that lies in the Pareto frontier, the new operating point for the algorithm is chosen as the solution with lower WER. This chapter also presented the reverberant speech database that is used in the implementation of the method, as well as the setup of the measures that are used. In the following chapters, the complete procedure will be applied to three different dereverberation algorithms.

Chapter 5

Enhancement of Dereverberation Algorithm A1

In this chapter the proposed method is first evaluated when applied to a one-microphone dereverberation algorithm, henceforth called A1. The quality and intelligibility of the speech signals processed by the original and enhanced algorithms are evaluated and compared, thus showing the degree of effectiveness of this work's proposal for this particular algorithm.

Dereverberation algorithm A1 was devised in [31] as a simplification of the two-stage dereverberation algorithm introduced by Wu and Wang [28]. The original two-stage algorithm consists of a first inverse-filter stage designed to reduce the coloration effect, and a second stage, based on spectral subtraction, designed for suppressing the long-term reverberation effect. In [31], results indicated an improvement in both quality and processing time by completely removing the first inverse-filtering stage.

The structure of the algorithm A1 and the methodology followed for its optimization, along with the experimental results obtained, are detailed in the following sections.

5.1 A1 description

The dereverberation algorithm A1, as depicted in Figure 5.1, aims at the estimation and subtraction of the long-term reverberation effect, which is caused by the late reverberation component of the RIR. This algorithm starts with the reverberant speech signal $z(n)$ and outputs the dereverberated speech $x(n)$, whose phase is determined directly from $z(n)$.

Let $S_z(m,k) = |S_z(m,k)|e^{j\varphi_z(m,k)}$ be the FFT of the n-th frame of the windowed version of $z(n)$, where a 32 ms Hamming window with 24 ms overlap between consecutive frames is used, and $w(m)$ be an asymmetrical smoothing window based on

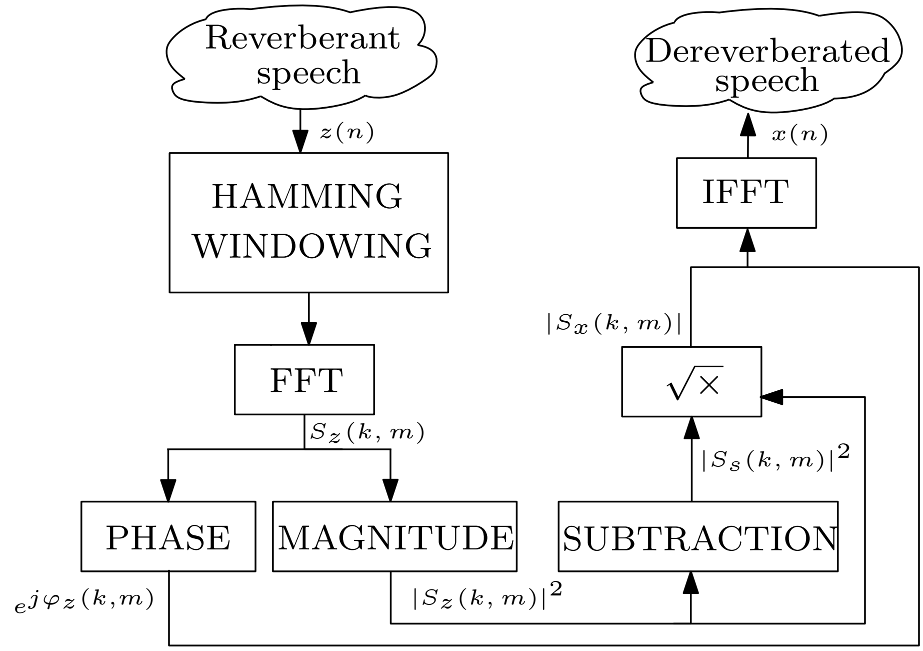


Figure 5.1: Diagram of the algorithm A1.

the Rayleigh distribution, given by

$$w(m) = \begin{cases} \left(\frac{m+a}{a^2}\right) e^{\left(\frac{-(m+a)^2}{2a^2}\right)}, & \text{if } m > -a \\ 0, & \text{otherwise} \end{cases}, \tag{5.1}$$

where parameter a controls the overall spread of the function.

The model of the power spectrum of the late reverberation can be described as

$$|S_l(m,k)|^2 = \gamma w(m-\rho) * |S_z(m,k)|^2, \tag{5.2}$$

where "$*$" represents the convolution operation along the time domain, k is the frequency bin, and m refers to the time frame. The parameter γ is a scaling factor and ρ represents the length of the early reflections.

Considering that the early and late components are mutually uncorrelated [28], the power spectrum of the early impulse components can be estimated by subtracting the power spectrum of the late impulse components from the reverberant speech. The spectrum subtraction scheme performs a weighting in the power spectrum of $z(n)$, and the block SUBTRACTION is given by

$$|S_s(m,k)|^2 = |S_z(m,k)|^2 \max\left[1 - \frac{|S_l(m,k)|^2}{|S_z(m,k)|^2}, \epsilon\right], \tag{5.3}$$

where ϵ is the floor and corresponds to the maximum attenuation. The power spectrum of $x(n)$ is given by

$$|S_x(m,k)|^2 = \sqrt{|S_z(m,k)|^2 \times |S_s(m,k)|^2}. \tag{5.4}$$

Finally, in order to calculate the spectrum of $x(n)$, the phase $\varphi_z(m,k)$ of $S_z(m,k)$ is combined to the magnitude $|S_x(m,k)|$, such that

$$S_x(m,k) = |S_x(m,k)|e^{j\varphi_z(m,k)}, \tag{5.5}$$

which allows one to estimate the clean signal $x(n)$ as desired.

5.2 Enhancement Process

This section describes the whole methodology used for the enhancement process of algorithm A1. The results and conclusions achieved here will serve as a basis for the enhancement of the remaining algorithms in this work. The key factor of the proposed method is focused on two aspects: the choice of the parameters under analysis and the objective-assessment measures employed. The proper management of these aspects facilitates the election of a new operating point for the algorithm, which can improve its performance.

5.2.1 Choice of the Parameters

For algorithm A1, four parameters were chosen to be optimized, as detailed below:

- Scaling factor (γ): Specifies the relative strength of the late-impulse components of the reverberant speech signal in Equation 5.2. Although many factors contribute to this relative strength (for instance, the reverberation time), the system performance is not very sensitive to specific values of γ [28]. The original value of the scaling factor in algorithm A1 was $\gamma = 0.35$.

- Attenuation limit (ϵ): Corresponds to the maximum attenuation in Equation 5.3. The original value of this parameter in algorithm A1 was $\epsilon = 0.001$, equivalent to an attenuation of 30 dB.

- Early-reflection length (ρ): Indicates the relative delay of the late impulse components in Equation 5.2. This delay reflects speech properties and is independent of reverberation characteristics. It is commonly considered to correspond to around 50 ms, which implies $\rho = 7$ frames. This value of ρ was set in algorithm A1.

- Spread control (a): This parameter controls the overall spread of function $w(n)$ in Equation 5.1. It needs to be less than or equal to ρ to provide a reasonable match to the equalized impulse-response shape. The original value of this parameter in algorithm A1 was $a = 6$.

These four parameters were combined within different ranges in order to proceed with the optimization strategy. Table 5.1 shows the parameter ranges considered for the enhancement of algorithm A1, which gives a total of 2475 training setups.

Table 5.1: Range of values of each parameter used in the enhancement process of algorithm A1.

Parameter	Range
γ	$\{0.30, 0.31, 0.32, \ldots, 0.40\}$
ϵ	$\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$
ρ	$\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$
a	$\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, with $a \leq \rho$

5.2.2 Choice of Objective-Assessment Measures

The combination of the parameters within their ranges were analyzed considering the average of different assessment measures, in the framework of the 12-signal training dataset. For this algorithm, four objective-assessment measures were combined and analyzed: SRMR, QAreverb, PESQ, and NCM, which were to be maximized. It is important to recall that both Q_{MOS} and PESQ measures are restricted to the 1–5 MOS range, whereas NCM values lies between 0 and 1.

Figures 5.2, 5.3, 5.4 and 5.5 show, respectively, the feasible objective space for the $Q_{MOS}\times$PESQ, SRMR×PESQ, NCM×PESQ, and NCM$\times Q_{MOS}$ relations, in which every scattered cross corresponds to one of the 2475 total combinations of the set $\{\gamma, \epsilon, \rho, a\}$. These figures also show the operating point of the original algorithm A1 [31] (labeled as "Original"), the point of the unprocessed reverberant signals (labeled as "Unprocessed"), and the operating point of the Wu and Wang algorithm [28] (labeled as "Wu-Wang"). On these figures, the Pareto-optimal solution is depicted. As can be observed, for all objective measures considered, the three labeled configurations of algorithm A1 lie in a solution which is dominated. A new solution among the Pareto-optimal points is expected to assume a better performance for the algorithm.

5.2.3 Choice of the Optimal Operating Point

Depending on the size and the shape of the objective feasible set, the number of Pareto-optimal solutions may be high. For example, on Figure 5.3 the cardinality

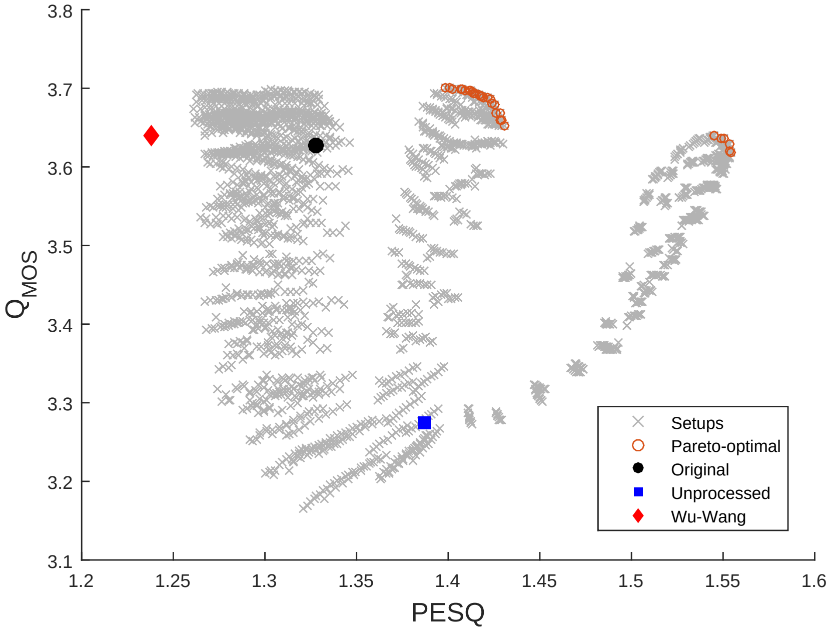


Figure 5.2: Q_{MOS}×PESQ relation for the training process of algorithm A1. Each scattered cross corresponds to a different combination of the set $\{\gamma, \epsilon, \rho, a\}$. This figure also shows the operating points corresponding to the original algorithm (circle), the unprocessed signals (square) and the Wu-Wang algorithm (diamond).

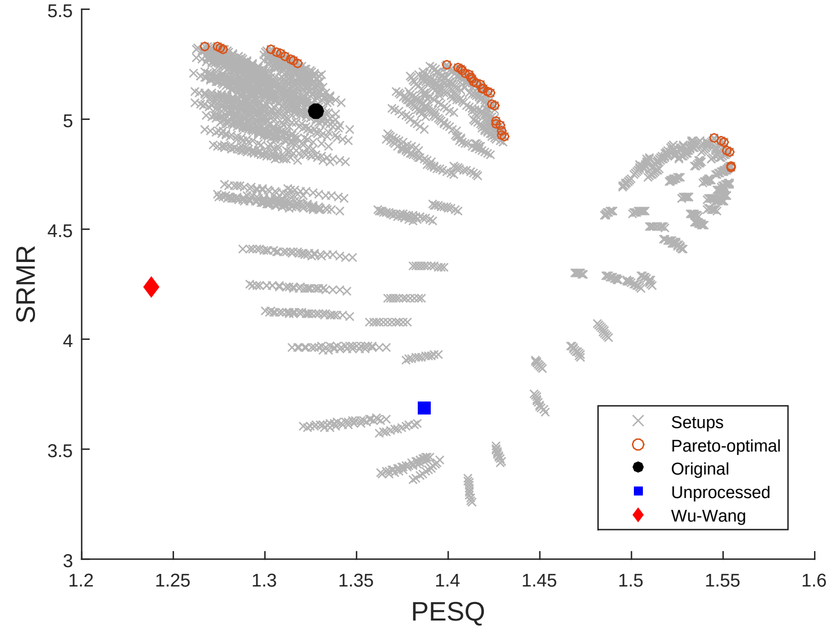


Figure 5.3: SRMR×PESQ graph for the training process of algorithm A1. Each scattered cross corresponds to a different combination of the set $\{\gamma, \epsilon, \rho, a\}$. This figure also shows the operating points corresponding to the original algorithm (circle), the unprocessed signals (square) and the Wu-Wang algorithm (diamond).

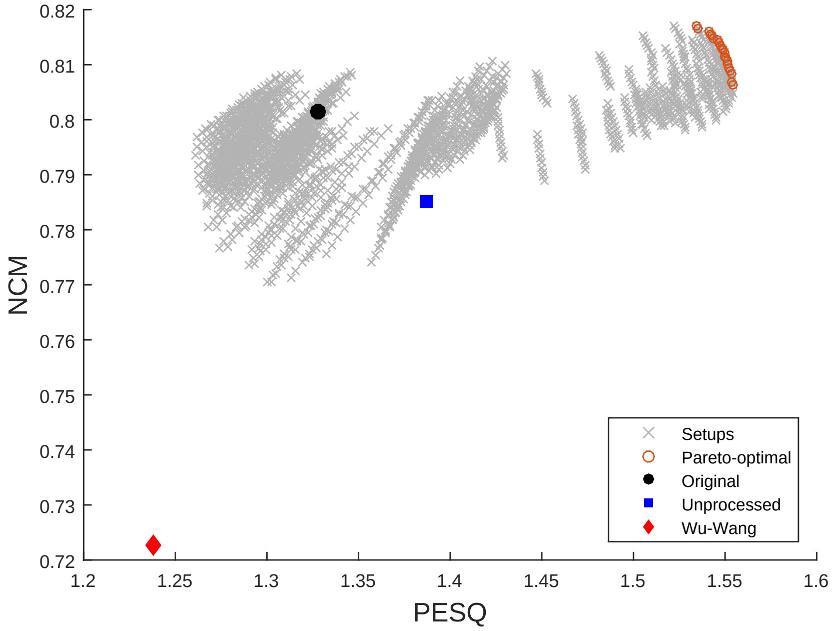


Figure 5.4: NCM×PESQ graph for the training process of algorithm A1. Each scattered cross corresponds to a different combination of the set $\{\gamma, \epsilon, \rho, a\}$. This figure also shows the operating points corresponding to the original algorithm (circle), the unprocessed signals (square) and the Wu-Wang algorithm (diamond).

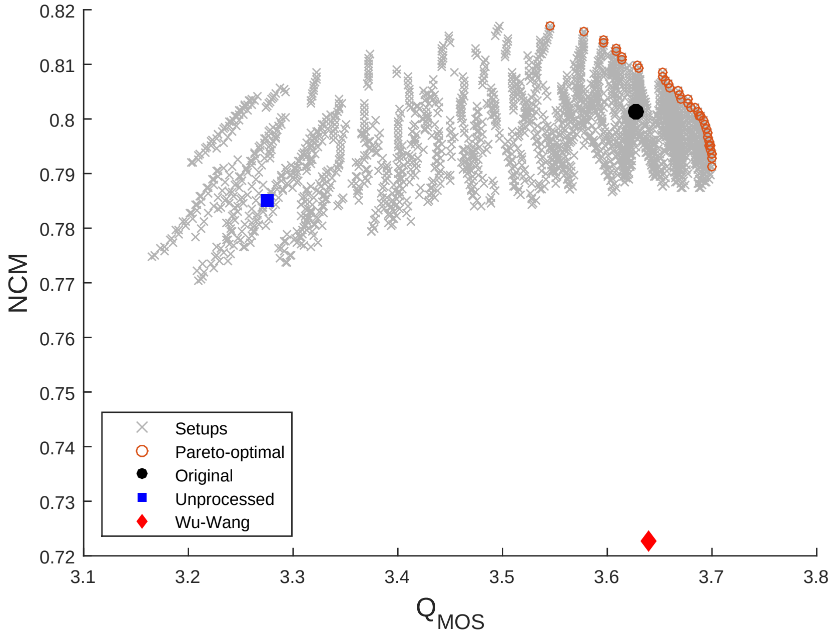


Figure 5.5: NCM×Q_{MOS} graph for the training process of algorithm A1. Each scattered cross corresponds to a different combination of the set $\{\gamma, \epsilon, \rho, a\}$. This figure also shows the operating points corresponding to the original algorithm (circle), the unprocessed signals (square) and the Wu-Wang algorithm (diamond).

of the Pareto-optimal set is 42. Analyzing all these solutions on the ASR system in order to compare the WER would be certainly an expensive task. The scope of this section is to evaluate the influence of the measures from the point of view of speech intelligibility, with the objective of reducing the search space for the optimal solutions. Different experiments are devised in order to establish a general rule to choose an optimal operating point for this and any other algorithm. In addition, it is determined which parameters most affect the performance of algorithm A1.

5.2.3.1 Experiment 1

The first experiment consists on determining the extent to which the increase of PESQ and Q_{MOS} measures affects the intelligibility of the speech signals. Figure 5.6 shows some selected operating points (from A to K) employed in this experiment. The average value of the WER was calculated for each one of these points, using the whole SimData and RealData development dataset. The points labeled as $\{A, B, C, D, E, F\}$ allow to evaluate the effect of increasing PESQ for two fixed values (medium and high) of Q_{MOS}. On the other hand, the points labeled as $\{G, D, H, I, J, K\}$ enable to evaluate the effect of increasing Q_{MOS} for two fixed values (medium and low) of PESQ. It is important to notice that, for this particular algorithm, a similar behavior between the two reverberation-based measures SRMR and Q_{MOS} is observed. Thus, the conclusions obtained here for the Q_{MOS} can be extended to the SRMR as well.

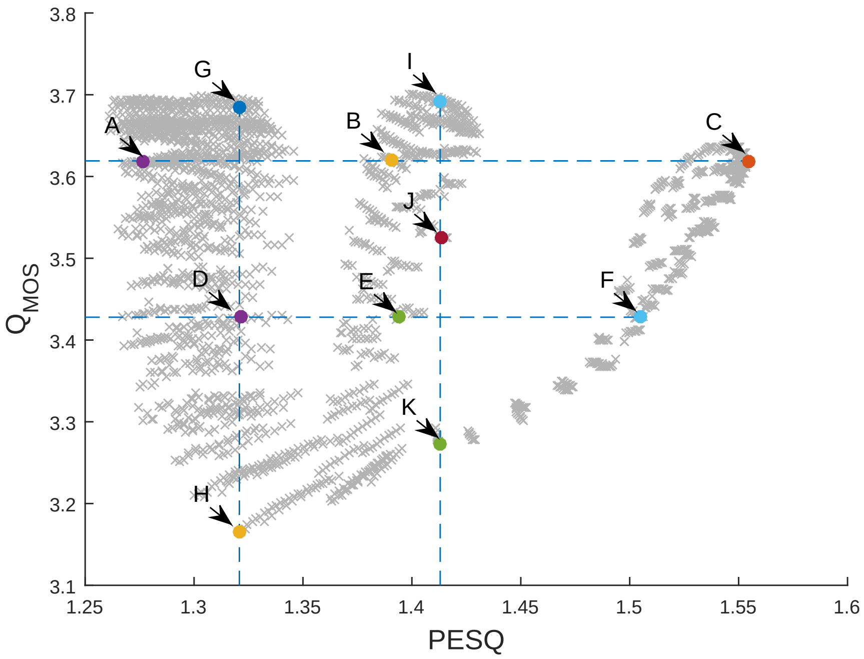


Figure 5.6: Operating points selected to evaluate the influence of PESQ and Q_{MOS} increase on speech intelligibility.

The analysis of the effect of increasing PESQ for fixed values of Q_{MOS} on the WER is depicted in Figure 5.7. It can be seen that, for medium and high values of Q_{MOS}, the increase of PESQ has a great impact on the reduction of the WER, thus improving speech intelligibility. On the other hand, Figure 5.8, which shows the effect of increasing Q_{MOS} for fixed values of PESQ, indicates that it is counterproductive to increase reverberant-based measures when PESQ values are low. This may be because the reduction of reverberation (and consequently the increase of measures like Q_{MOS} and SRMR) using algorithm A1 can introduce artifacts that affect speech intelligibility. By including a more general purpose quality measure, such as PESQ, reverberation can be reduced with algorithm A1 more effectively without introducing these artifacts. Indeed, as shown in Figure 5.8, for medium values of PESQ, the increase in Q_{MOS} reduces the WER. On this curve, it is also possible to appreciate a saturation profile starting from point J, which means that there comes a moment in which reducing reverberation with algorithm A1 might not improve intelligibility.

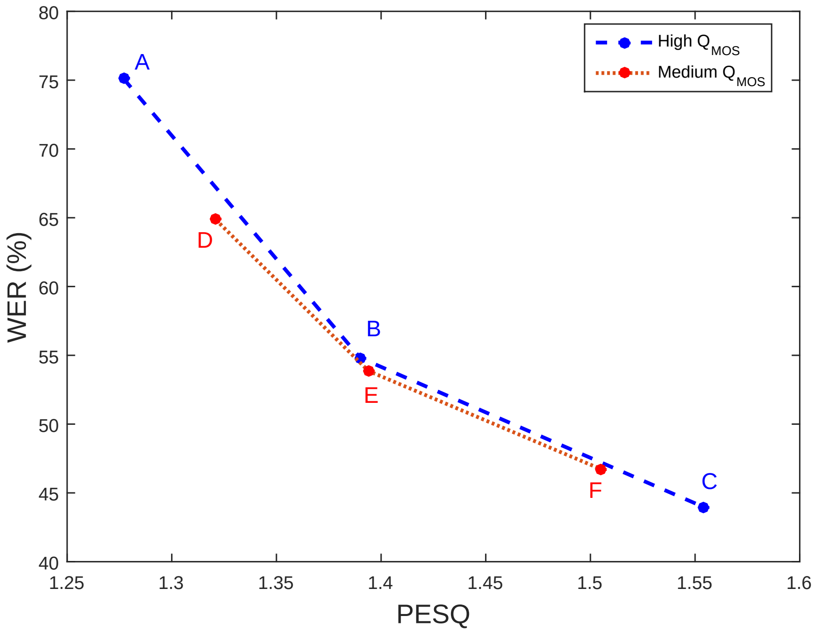


Figure 5.7: WER×PESQ relation for different fixed values of Q_{MOS}.

To give the reader a better perspective of the results of this experiment, Figure 5.9 shows the average of the WER for each one of the selected operating points directly in the Q_{MOS}×PESQ plot.

On the basis of the results of this experiment, it can be determined that the preferred solutions among the Pareto-optimal points lies in the region of higher PESQ values.

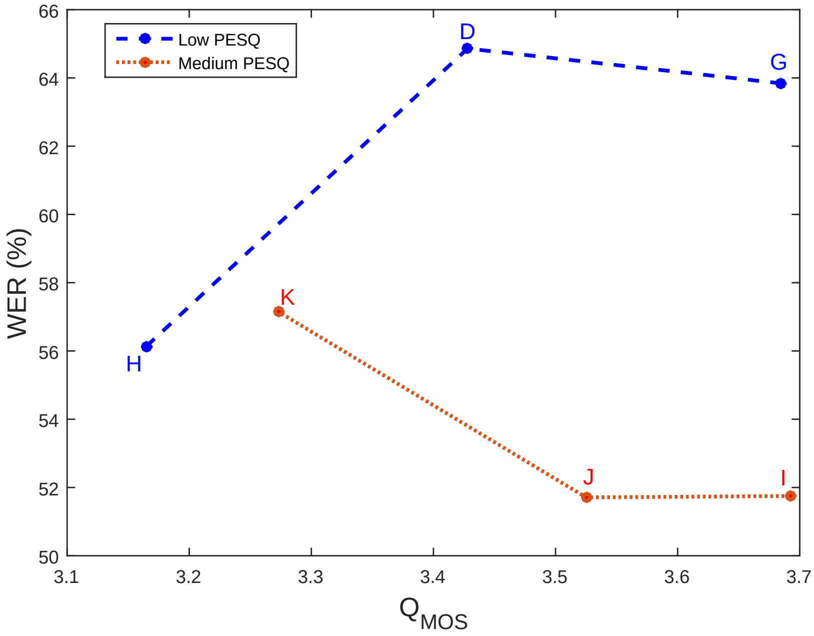


Figure 5.8: WER×Q_{MOS} relation for different fixed values of PESQ.

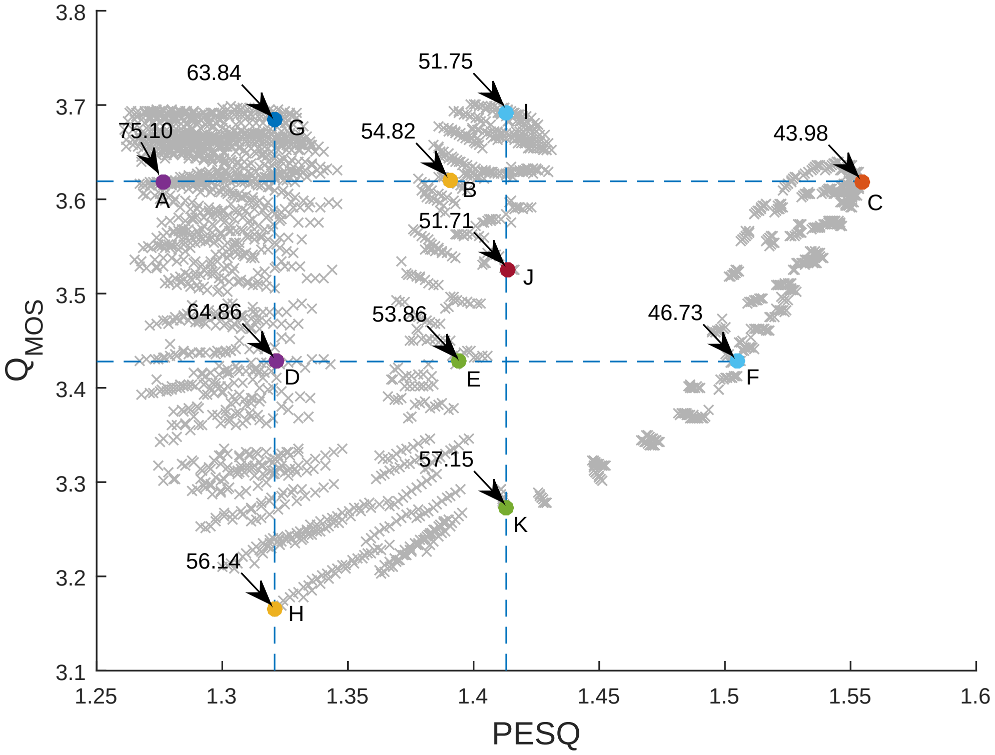


Figure 5.9: Average of WER for each one of the selected points on the Q_{MOS}×PESQ plot.

5.2.3.2 Experiment 2

Having concluded that the joint increase in both Q_{MOS} and PESQ improves speech intelligibility, this experiment aims at choosing the optimal operating point considering these two measures simultaneously. For this purpose, some non-dominated points from the Pareto-optimal set in which Q_{MOS} and PESQ are jointly maximized were analyzed, as shown in Figure 5.10. Within this reduced set, five operating points of the Pareto frontier (from $H1$ to $H5$) were chosen in order to compare their performances with respect to speech intelligibility. It is worth mentioning that an equivalent frontier would be found by replacing Q_{MOS} for SRMR.

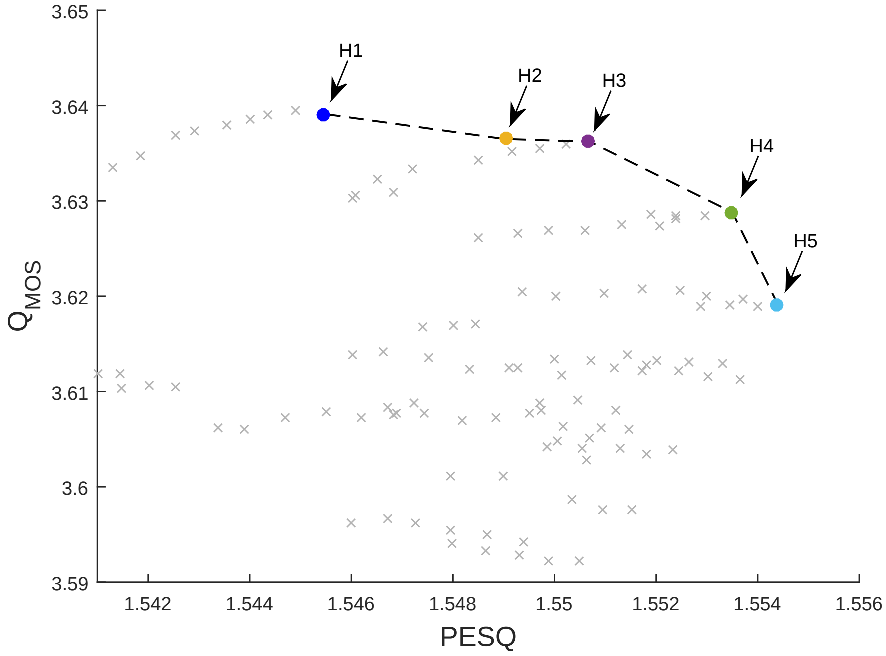


Figure 5.10: Some non-dominated solutions located on the Pareto frontier that jointly maximizes the PESQ and Q_{MOS} measures.

Figure 5.11 shows the average of the WER for points $H1$ to $H5$ using the whole SimData and RealData development dataset. The best result is obtained for the point $H3$, which corresponds to the ‘elbow’ of the frontier. This point is considered the new optimal operating point of the enhanced algorithm A1, corresponding to the set $\{\gamma = 0.39, \epsilon = 0.1, \rho = 9, a = 5\}$.

5.2.3.3 Experiment 3

For this experiment, the effect of the NCM was analyzed. As stated in Chapter 3, this measure had been shown to correlate highly with the intelligibility of reverberant speech in other works [61]. To carry on this experiment, several points located on the Pareto-frontier that jointly maximize the NCM and PESQ were analyzed,

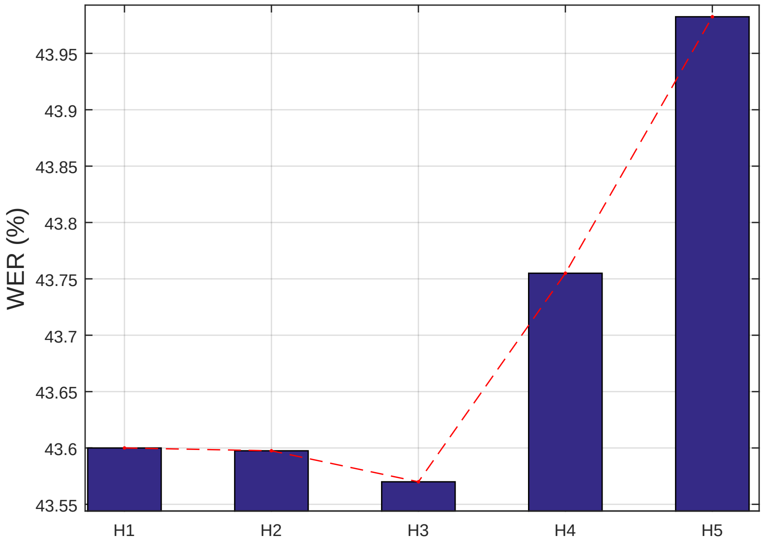


Figure 5.11: Average of WER for candidates to optimal operating point considering the PESQ and Q_{MOS} measures.

as shown in Figure 5.12, in which points $H1$ to $H5$ are also included. As can be observed, points $H1$ to $H4$ are not Pareto-optimal solutions for this particular feasible objective space.

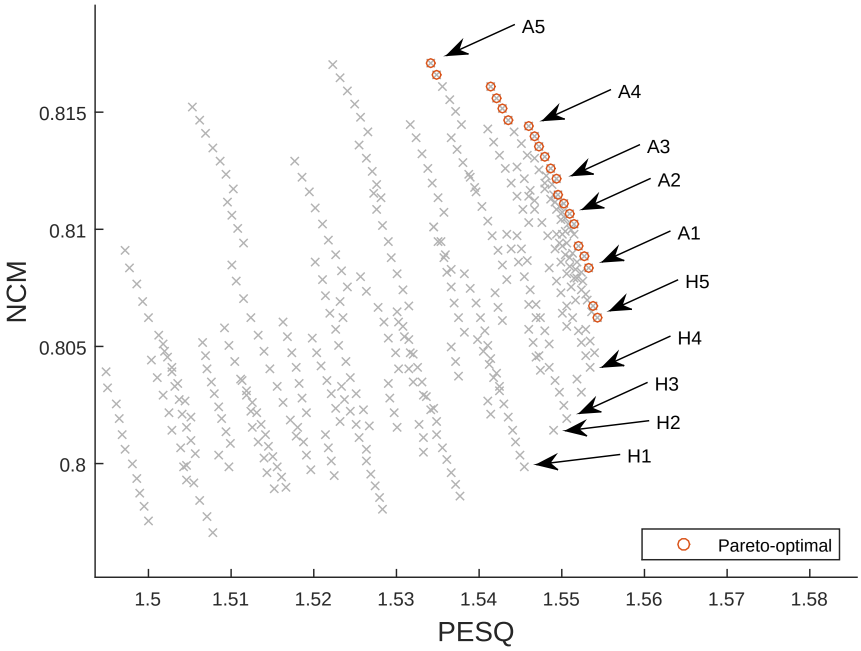


Figure 5.12: Operating points located on the convex hull that jointly maximizes the NCM and PESQ values.

As shown in Figure 5.13, results show that the WER is not improved by increasing the NCM.

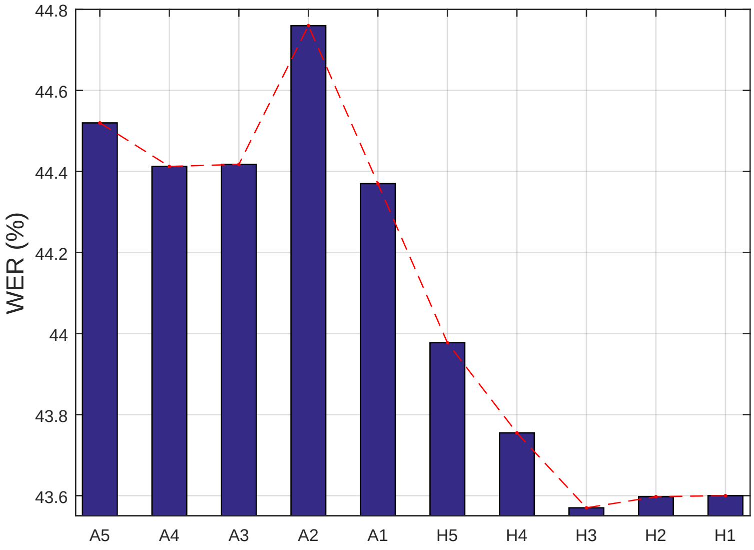


Figure 5.13: Average of WER for candidates to optimal operating point considering the NCM and PESQ values.

5.3 Overall Results

Four of the parameters in algorithm A1 (namely, the scaling factor γ, attenuation limit ϵ, length of early reflections ρ, and spread control a) were finely tuned following the optimization method proposed in this work. The best solution was found by using the Q_{MOS}, the SRMR and the PESQ measures as objectives to be optimized. Figure 5.14 shows the $Q_{MOS}\times$PESQ and SRMR×PESQ relations for the training database, marking the original and optimized operating points (corresponding to the sets $\{\gamma = 0.35, \epsilon = 0.001, \rho = 7, a = 6\}$ and $\{\gamma = 0.39, \epsilon = 0.1, \rho = 9, a = 5\}$, respectively), as well as the point corresponding to unprocessed reverberant signals.

For both development and evaluation datasets, Tables 5.2 and 5.3 compare, respectively, the PESQ, Q_{MOS}, SRMR, and WER measures of the unprocessed reverberant signals (Unprocessed scenario), the signals processed by the unmodified algorithm A1 [31] (Original scenario), and the signals processed by the optimized algorithm (Optimized scenario). Regarding quality measures, the results of these tables fit with the results shown on Figure 5.14, that is, the optimized operating point has higher PESQ, similar Q_{MOS} and lower SRMR in relation to the original point in almost all rooms and distances for the training and the whole SimData datasets. This shows that the training signals generalize well for the complete database.

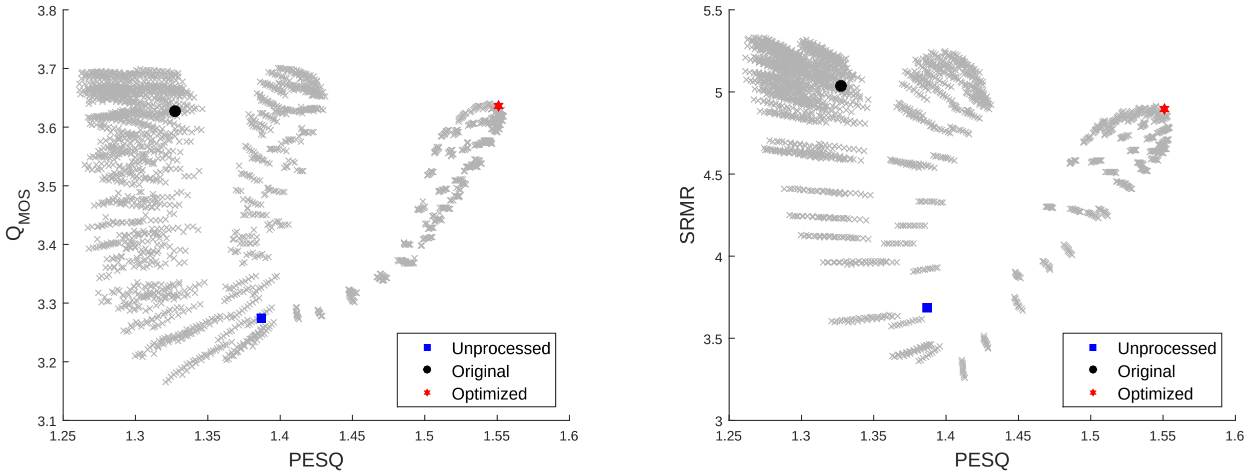


Figure 5.14: Q_{MOS}×PESQ (left) and SRMR×PESQ (right) plots for the training database, showing the unprocessed, original, and optimized operating points for algorithm A1.

In relation to the WER values, the optimized scenario presents a better performance for all rooms except for the first one, in which the unprocessed signals get slightly better results. In comparison to the original algorithm configuration, the optimized scenario presents an average improvement of 24% for SimData and 10% for RealData, thus veryfying the effectiveness of the proposed method from the point of view of speech intelligibility.

For the RealData set, due to the lack of the anechoic signals, it is only possible to use blind metric approaches, such as the SRMR and the blind version of Q_{MOS}. Although these measure values are higher in the case of the original algorithm configuration, WER values are lower for the optimized scenario, thus showing that a good score in reverberant based metrics does not always imply an improvement in speech intelligibility.

Table 5.2: Results of algorithm A1 for SimData and RealData development datasets. Bold numbers indicate the best results.

		SimData							RealData		
Measure	Scenario	Room 1		Room 2		Room 3		Avg.	Room 4		Avg.
		Near	Far	Near	Far	Near	Far	–	Near	Far	–
PESQ	Unprocessed	2.09	1.35	1.39	1.16	1.36	1.16	1.42	-	-	-
	Original	1.53	1.28	1.43	1.25	1.42	1.26	1.36	-	-	-
	Optimized	**2.20**	**1.44**	**1.67**	**1.27**	**1.64**	**1.26**	**1.58**	-	-	-
Q_{MOS}	Unprocessed	**4.23**	3.90	3.52	2.33	3.27	2.38	3.27	2.46	2.41	2.43
	Original	3.86	3.95	3.77	**3.07**	3.63	**3.19**	3.58	**7.92**	**7.15**	**7.54**
	Optimized	4.11	**3.97**	**3.82**	2.96	**3.66**	3.09	**3.60**	3.31	3.27	3.29
SRMR	Unprocessed	4.37	4.63	3.67	2.94	3.66	2.76	3.67	4.06	3.52	3.79
	Original	5.05	**5.82**	**4.86**	**4.97**	**4.94**	**4.82**	**5.08**	**7.92**	**7.15**	**7.54**
	Optimized	**5.10**	5.72	4.76	4.70	4.81	4.54	4.94	7.27	6.55	6.91
WER (%)	Unprocessed	**15.29**	**25.29**	43.90	85.80	51.95	88.9	51.81	88.71	88.31	88.51
	Original	53.98	64.58	51.00	66.90	59.94	69.76	61.02	73.61	74.16	73.88
	Optimized	18.71	25.84	**26.82**	**57.48**	**33.09**	**60.56**	**37.06**	**61.63**	**64.59**	**63.10**

Table 5.3: Results of algorithm A1 for SimData and RealData evaluation datasets. Bold numbers indicate the best results.

		SimData							RealData		
Measure	Scenario	Room 1		Room 2		Room 3		Avg.	Room 4		Avg.
		Near	Far	Near	Far	Near	Far	–	Near	Far	–
PESQ	Unprocessed	2.14	1.60	1.40	1.19	1.37	1.17	1.48	-	-	-
	Original	1.59	1.46	1.49	1.30	1.51	1.27	1.44	-	-	-
	Optimized	**2.25**	**1.72**	**1.75**	**1.32**	**1.68**	**1.28**	**1.67**	-	-	-
Q_{MOS}	Unprocessed	**4.24**	3.96	3.61	2.38	3.20	2.40	3.30	2.51	2.57	2.54
	Original	3.97	**4.03**	3.87	**3.16**	**3.71**	**3.22**	**3.66**	**4.22**	**4.21**	**4.22**
	Optimized	4.16	4.03	**3.90**	3.01	3.68	3.10	3.65	3.35	3.37	3.36
SRMR	Unprocessed	4.50	4.58	3.74	2.97	3.57	2.73	3.68	3.17	3.19	3.18
	Original	5.25	**5.75**	**5.13**	**5.03**	**5.11**	**4.82**	**5.18**	**6.20**	**6.37**	**6.28**
	Optimized	**5.28**	5.70	4.99	4.73	4.92	4.51	5.02	5.66	5.81	5.74
WER (%)	Unprocessed	**18.06**	**25.38**	42.98	82.20	53.54	88.04	51.68	89.72	87.34	88.53
	Original	61.28	68.50	49.22	62.12	58.76	72.27	62.02	81.48	79.64	80.56
	Optimized	23.40	28.48	**27.13**	**50.73**	**35.81**	**62.00**	**37.91**	**72.37**	**69.21**	**70.79**

Figures 5.15, 5.16, and 5.17 show comparative bar graphs for the WER, Q_{MOS}, SRMR, and PESQ for all databases. The graphs of SRMR, Q_{MOS}, and PESQ also include the confidence intervals. These figures may offer a better perspective of the results obtained. For example, Figure 5.15 shows how the WER of the optimized scenario is improved in relation to the original configuration for every room and source-microphone distance. Figures 5.16 and 5.17 shows how the SRMR and Q_{MOS} measures take similar values in the original and the optimized scenarios, whereas the optimized configuration has a higher PESQ value for every room and distance.

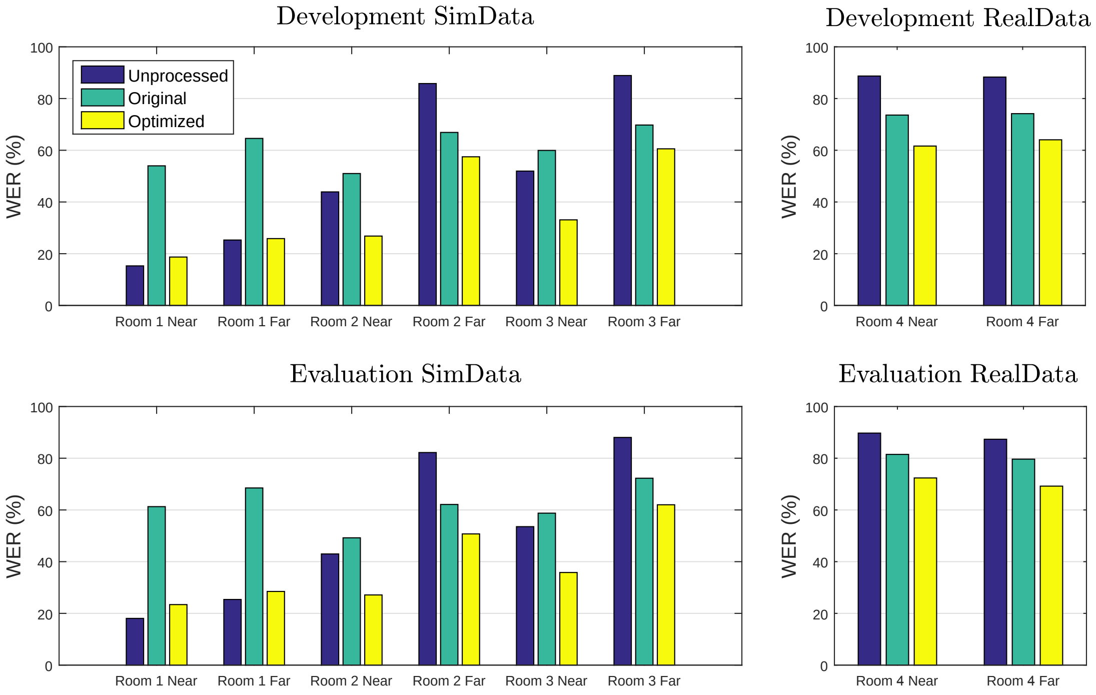


Figure 5.15: WER results for the whole database comparing the different configurations of algorithm A1.

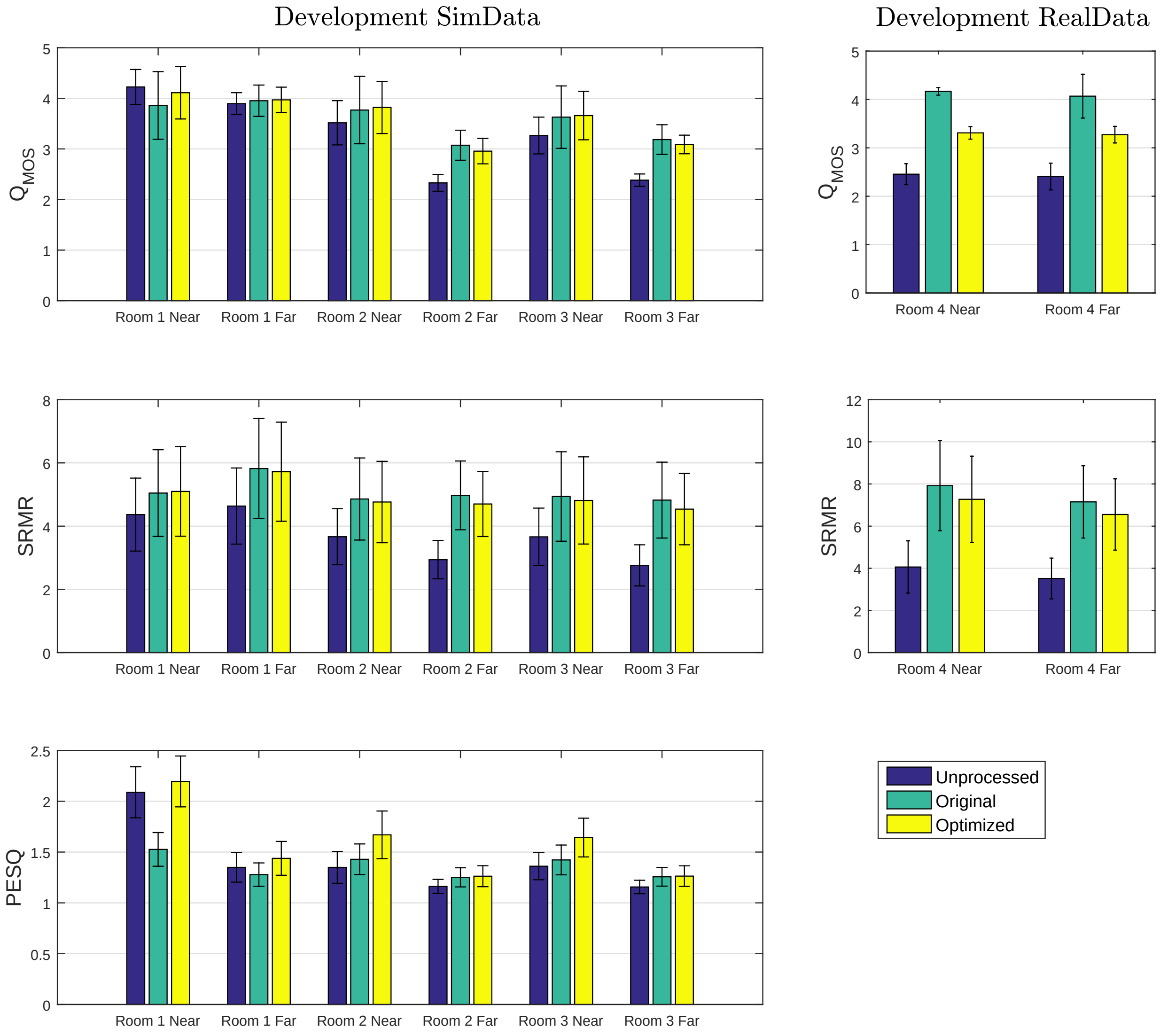


Figure 5.16: Q_{MOS}, SRMR and PESQ results for SimData and RealData development dataset comparing the different configurations of algorithm A1.

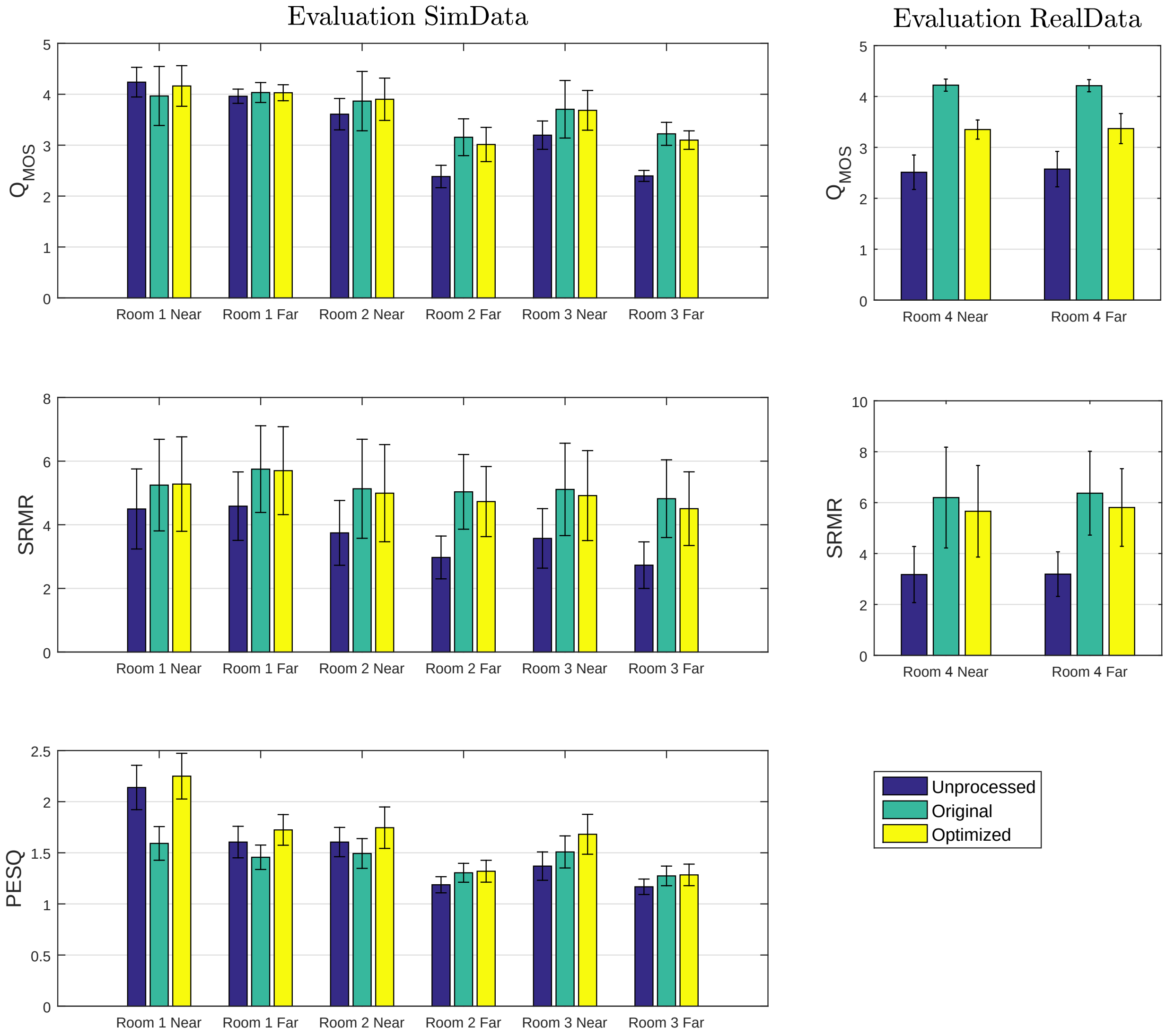


Figure 5.17: Q_{MOS}, SRMR and PESQ results for SimData and RealData evaluation dataset comparing the different configurations of algorithm A1.

5.4 Conclusion

Algorithm A1, based on a single-channel blind spectral subtraction, was the first algorithm to be enhanced according to the proposed methodology. Three experiments were carried out in the process of optimization. The main findings are summarized below:

- The PESQ measure is highly correlated with speech intelligibility.
- When the PESQ is low, it is counterproductive to increase Q_{MOS} (or SRMR) in terms of intelligibility.
- The best operating point (that is, the point of minimum WER) for the algorithm is found at the 'elbow' of the Pareto frontier that jointly maximizes PESQ and Q_{MOS} (or SRMR) measures.
- The NCM does not have a significant correlation with speech intelligibility.

Table 5.4 compares the original [31] and enhanced configuration of algorithm A1 in terms of parameter values and total average WER achieved by the ASR system. Results demonstrate the effectiveness of the proposed approach as it led to an algorithm scenario that outperformed the original configuration in terms of speech intelligibility, as assessed by the improvement of the WER (by an average of 22%) achieved by the ASR system.

Table 5.4: Comparison of the parameter values and the total average WER between the original and the enhanced configuration of algorithm A1.

	Original Configuration	Enhanced Configuration
Parameters	$\gamma = 0.35$ $\epsilon = 0.001$ $\rho = 7$ $a = 6$	$\gamma = 0.39$ $\epsilon = 0.1$ $\rho = 9$ $a = 5$
WER(%)	63.81	41.59

Chapter 6

Enhancement of Dereverberation Algorithm A2

This chapter presents the enhancement process of another single-channel algorithm, henceforth A2, that is based on a statistical model of reverberation. This algorithm suppresses both noise and late reverberation through the application of gain coefficients in a time-frequency domain. Algorithm A2 is more complex than the algorithm A1 (analyzed in the previous chapter), and involves the tuning of a larger number of parameters. The baseline algorithm, as proposed in [21], uses three different configurations that differ in the time-frequency implementation. The enhanced version of algorithm A2 is compared with these three configurations in terms of quality measures and WER.

The structure of algorithm A2 and the methodology followed for its optimization, along with the experimental results obtained, are detailed in the following sections.

6.1 A2 Description

This single-microphone spectral enhancement algorithm, proposed in [21], is able to suppress both late reverberation and background noise using statistical models for the reverberation process. Its structure, as depicted in Figure 6.1, revolves around the minimum mean-square error log-spectral amplitude (MMSE-LSA) enhancement stage, which is detailed next. Regarding the time-frequency analysis, the authors proposed two different transforms: the traditional STFT and the short-time fan-chirp transform (STFChT), which allows one to use longer analysis windows [68]. Due to the complexity of algorithm A2, the steps of this complete algorithm are summarized at the end of this section.

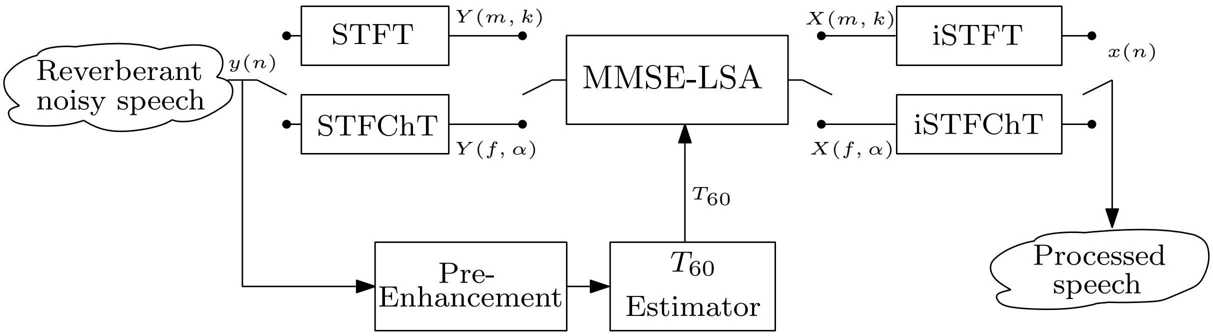


Figure 6.1: Diagram of algorithm A2.

6.1.1 Minimum Mean Square Error Log Spectral Amplitude estimator

This speech enhancement method was originally proposed in [69] and later improved in [70]. Originally, this method was focused on noise reduction. Later, Habets [3] proposed some modifications to suppress both noise and late reverberation. With these modifications the RIR is considered to be partitioned into two components (early and late), such that

$$h(n) = \begin{cases} 0, & n < 0 \\ h_e(n), & 0 \leq n < n_e \\ h_l(n), & n > n_e \end{cases}, \tag{6.1}$$

where n_e is the number of samples that constitute the direct-path and a few early reflections of the RIR, as shown in Figure 6.2.

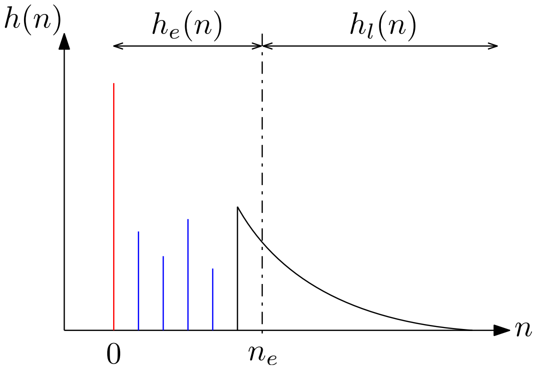


Figure 6.2: Schematic representation of the RIR division into early and late components.

The received signal, $y(n)$, at the microphone in a noisy and reverberant environment can be then modeled as

$$y(n) = \underbrace{z_e(n) + z_l(n)}_{z(n)} + \nu(n), \tag{6.2}$$

where $z_e(n)$ and $z_l(n)$ correspond to the early (including the direct-path) and late reverberant speech signals, respectively. Term $\nu(n)$ corresponds to the additive noise. At the same time, term $z(n)$ can be expressed as

$$z(n) = z_d(n) + z_r(n), \tag{6.3}$$

where subscripts d and r stand for the direct and reverberant components of the speech, respectively.

The objective of this algorithm is to jointly suppress the terms $z_l(n)$ and $\nu(n)$ from Equation 6.2 in order to reduce the effective noise level as well as to increase speech fidelity and intelligibility, while maintaining the colorations caused by early reflections. Thus, the MMSE-LSA estimator yields an estimate of $z_e(n)$ in the time-frequency domain, $\hat{Z}_e(m,k)$, without using detailed knowledge of the RIR, by applying a frequency-dependent gain $G_{MMSE-LSA}(n,k)$ to the noisy and reverberant spectral coefficient $Y(m,k)$:

$$\hat{Z}_e(m,k) = G_{MMSE-LSA}(m,k)Y(m,k). \tag{6.4}$$

The MMSE-LSA gain function is computed as

$$G_{MMSE-LSA}(m,k) = G_{LSA}(m,k)^{p(m,k)} G_{min}(m,k)^{1-p(m,k)}, \tag{6.5}$$

where $p(m,k)$ is the probability that the desired speech component $z_e(n)$ is present, and $G_{LSA}(m,k)$ and $G_{min}(m,k)$ are given by

$$G_{LSA}(m,k) = \frac{\xi(m,k)}{1+\xi(m,k)} \exp\left(\frac{1}{2}\int_{\varsigma(m,k)}^{\infty} \frac{e^{-t}}{t} dt\right) \tag{6.6}$$

$$G_{min}(m,k) = \frac{G_{min,z_l}\hat{\lambda}_{z_l}(m,k) + G_{min,\nu}\hat{\lambda}_{v}(m,k)}{\hat{\lambda}_{z_l}(m,k) + \hat{\lambda}_{\nu}(m,k)}. \tag{6.7}$$

All terms involved in Equations 6.6 and 6.7 are described below:

- $\xi(m,k)$ denotes the *a priori* signal-to-interference ratio (SIR), computed as

$$\xi(m,k) = \frac{\lambda_{z_e}(m,k)}{\lambda_{z_l}(m,k) + \lambda_{z_\nu}(m,k)}, \tag{6.8}$$

where terms $\lambda_{z_e}(m,k), \lambda_{z_l}(m,k)$ and $\lambda_{z_\nu}(m,k)$ are the spectral variances[1] of the early speech component, late reverberation, and ambient noise, respectively.

- $\varsigma(m,k) = \frac{\xi(m,k)}{1+\xi(m,k)}\gamma(m,k)$, where $\gamma(m,k)$ denote the *a posteriori* SIR, given by
$$\gamma(m,k) = \frac{|Y(m,k)|^2}{\lambda_{z_l}(m,k) + \lambda_\nu(m,k)}. \tag{6.9}$$

- $\hat{\lambda}_{z_l}(m,k)$ and $\hat{\lambda}_\nu(m,k)$ are the estimated late reverberant spectral variance and the estimated ambient noise spectral variance, respectively.

- G_{min,z_l} and $G_{min,\nu}$ are used to control the maximum suppression of late reverberation and ambient noise, respectively. When $G_{min,z_l} = 0$, the late reverberation is suppressed down to the residual level of ambient noise.

Therefore, in order to compute the MMSE-LSA estimator, it is necessary to estimate the *a priori* SIR, the late reverberant spectral variance $\hat{\lambda}_{z_l}(m,k)$, and the ambient noise spectral variance $\hat{\lambda}_\nu(m,k)$. All these estimators are detailed next.

6.1.1.1 *A priori* SIR Estimator

First define the *a priori* SIR of each interference (noise and reverberation) separately, that is, $\xi_{z_l}(m,k) = \frac{\lambda_{z_e}(m,k)}{\lambda_{z_l}(m,k)}$ and $\xi_\nu(m,k) = \frac{\lambda_{z_e}(m,k)}{\lambda_\nu(m,k)}$. The *a priori* SIR estimation can be calculated as follows [3]:

$$\xi(m,k) = \left\{ \begin{array}{ll} \xi_v(m,k) & 10\log_{10}\left(\frac{\lambda_\nu(m,k)}{\lambda_{z_l}(m,k)}\right) > \beta^{dB} \\ \frac{\xi_{z_l}(m,k)\xi_\nu(m,k)}{\xi_{z_l}(m,k)+\xi_\nu(m,k)} & \text{otherwise} \end{array} \right. , \tag{6.10}$$

where the threshold β^{dB} specifies the level difference between $\lambda_\nu(m,k)$ and $\lambda_{z_l}(m,k)$ in dB. To estimate $\xi_\vartheta(m,k)$, where $\vartheta \in \{z_l, \nu\}$, the following expression is used [69]:

$$\hat{\xi}_\vartheta(m,k) = \max\left\{\eta_\vartheta \frac{G^2_{LSA}(n-1,k)|Y(n-1,k)|^2}{\lambda_\vartheta(n-1,k)} + (1-\eta_\vartheta)\Psi_\vartheta(m,k), \xi_{min,\vartheta}\right\}, \tag{6.11}$$

where

$$\Psi_\vartheta(m,k) = \frac{\lambda_{z_l}(m,k) + \lambda_\nu(m,k)}{\lambda_\vartheta(m,k)}[\gamma(m,k) - 1], \tag{6.12}$$

and $\xi_{min,\vartheta}$ is a lower bound.

[1] The spectral variance of a variable x is computed as $\lambda_x(m,k) = E\{|X(m,k)|^2\}$

6.1.1.2 Estimator for the Ambient Noise Spectral Variance

A statistical-model-based voice activity detector (VAD) is used to update the noise spectral variance during speech-absent periods. This method, proposed in [4], compares the following quantity to a threshold η_{thresh}:

$$\eta(n) = \sum_k \ln \gamma(m,k) \frac{\xi(m,k)}{1+\xi(m,k)} - \ln(1+\xi(m,k)). \tag{6.13}$$

If $\eta(n) < \eta_{thresh}$, the frame is considered to be only noise and the noise variance is updated as follows:

$$\lambda_v(m,k) = \mu_\nu \lambda_\nu(n-1,k) + (1-\mu_\nu)|Y(m,k)|^2 \tag{6.14}$$

6.1.1.3 Estimator for the Late Reverberant Spectral Variance

The spectral variance of the late reverberant component can be expressed as [3]:

$$\lambda_{z_l}(m,k) = \mathrm{e}^{\bar{\zeta}(k)(n_e-R)} \lambda_{z_r}(n - \frac{n_e}{R} + 1, k), \tag{6.15}$$

where:

- $\bar{\zeta}(k)$ is related to the reverberation time by

  $$\bar{\zeta}(k) = \frac{3\ln(10)}{T_{60}(k) f_s}, \tag{6.16}$$

  where f_s denotes the sampling frequency;

- R is the number of samples separating two successive analysis frames;

- $\lambda_{z_r}(m,k)$ is the spectral variance of the reverberant component of the speech signal (that is, disregarding the direct path), computed as

  $$\lambda_{z_r}(m,k) = \mathrm{e}^{-2\bar{\zeta}(k)R}(1-\kappa(k))\lambda_{z_r}(n-1,k) + \kappa(k)\mathrm{e}^{-2\bar{\zeta}(k)R}\lambda_z(n-1,k), \tag{6.17}$$

  where parameter κ is related to the inverse of the DRR, $\kappa \propto \frac{E_r}{E_d}$

Therefore, the variance of the late-reverberant component can be computed from the variance of the total reverberant component. It is important to notice that for this calculation a blind estimation of the DRR and T_{60} values is also required. For T_{60} estimation, algorithm A2 uses the method proposed by Löllmann et al. [15]. For the DRR estimation, it is used an online adaptive procedure [3], since the DRR depends on the microphone-source distance.

Algorithm A2 Summary of the single-microphone spectral enhancement algorithm A2, that suppresses late reverberation and ambient noise.

```
**Pre-enhancement:** First, a noise reduction algorithm based on LSA [69] is performed.
$T_{60}$ **estimation:** Using the blind algorithm proposed in [15].
**STFT or STFChT**: Calculate the STFT or STFChT of the signal.
**MMSE-LSA**: Compute the MMSE-LSA estimator as follows:
    i Estimate model parameters: $\bar{\zeta}(k)$ using (6.16) and $\kappa$ according to [3].
    ii Estimate ambient noise: Assuming that the first 6 frames are noise, the spectral variance of the noise is actualized online as in Section 6.1.1.2.
    iii Estimate late reverberant energy $\hat{\lambda}_{z_l}(m,k)$ using (6.15).
    iv Calculate the *a posteriori* SIR using (6.9), the individual *a priori* SIR using (6.11), and the total *a priori* SIR using (6.10).
    v Estimate the speech presence probability $p(m,k)$ using the method described in [71].
    vi Calculate the gain function $G_{MMSE-LSA}(m,k)$ using (6.6), (6.7), and (6.5).
    vii Calculate $\hat{Z}_e(m,k)$ using (6.4).
**Inverse STFT or STFChT**: Calculate the output $x(n) = \hat{z}_e(n)$ by applying the inverse STFT or STFChT, as appropriate, to $\hat{Z}_e(m,k)$.
```

6.2 Enhancement Process

Three configurations of algorithm A2 (named here as A2.1, A2.2 and A2.3) were proposed in [21], as detailed below. All these versions employ a Hamming analysis window and a frame hop of 128 samples. For the enhancement process, the STFT was chosen for time-frequency analysis.

- A2.1: Uses the STFT with a short window (512-sample long) and an FFT length of 512.
- A2.2: Uses the STFT with a long window (2048-sample long) and an FFT of 3262.
- A2.3: Uses the STFChT with a long window and an FFT length of 3262.

The complete procedure for the optimization of this algorithm is described next.

6.2.1 Choice of the parameters

Among the parameters in algorithm A2, twelve of them were selected for the enhancement process. Since the original configurations of this algorithm employ different FFT and analysis window lengths, these parameters are also included in the

Table 6.1: Parameters chosen for the enhancement process of algorithm A2, including a brief description, the equation in which they are involved, their original values, and the range of analysis.

Param.	Description	Eq.	Original value	Range
N_{FFT}	Length of the FFT in samples.	–	–	$\{512, 1024, 2048\}$
N_{win}	Length of the analysis Hamming window in samples. It has to be less than or equal to the length of the FFT.	–	–	$\{512, 1024, 2048\}$
n_e	Number of samples that constitute the direct-path and a few early reflections.	(6.15)	768 (∼50 ms)	$\{640, 768, 896, 1024\}$
G_{min,z_l}	Controls the maximum suppression of late reverberation.	(6.7)	0 dB	$\{-5, 0, 5\}$
$G_{min,\nu}$	Controls the maximum suppression of ambient noise.	(6.7)	-12 dB	$\{-15, -10, -5, 0\}$
β^{dB}	Threshold that specifies the level difference between $\lambda_\nu(m, k)$ and $\lambda_{z_l}(m, k)$.	(6.10)	3 dB	$\{2, 3, 4\}$
η_{z_l}	Weighting factor that controls the tradeoff between the amount of noise reduction and distortion on the a priori SIR of the late reverberant interference.	(6.11)	0.95	$\{0.85, 0.86, 0.87, \ldots, 0.95\}$
η_ν	Weighting factor that controls the tradeoff between the amount of noise reduction and distortion on the a priori SIR of the noise interference.	(6.11)	0.95	$\{0.85, 0.86, 0.87, \ldots, 0.95\}$
ξ_{min,z_l}	Lower bound on the a priori SIR of the late reverberant interference.	(6.11)	-25 dB	$\{-30, -25, -20, -15\}$
$\xi_{min,\nu}$	Lower bound on the a priori SIR of the noise interference.	(6.11)	-25 dB	$\{-30, -25, -20, -15\}$
η_{thresh}	Threshold used in the voice activity detector for the update of the noise variance.	(6.13)	0.3	$\{0.1, 0.2, 0.3, 0.4\}$
μ_ν	Weighting factor for the update of the noise variance.	(6.14)	0.98	$\{0.85, 0.86, 0.87, \ldots, 0.99\}$

optimization process. Table 6.1 summarizes the selected parameters, including their related equation, original values, and range of analysis for the optimization process.

For computational reasons, it was not possible to combine every parameter value with each other for this particular algorithm. Thus, the parameters were divided into three groups, as shown in Table 6.2, which also shows the number of combinations of parameter values for each group. First, the best combination of the FFT and window analysis lengths was evaluated. The rest of the considered parameters was divided between the second and third groups. In total, 8665 different configurations of algorithm A2 were evaluated.

In order to proceed with the optimization methodology, the optimal parameters selected for the first group are fixed in the optimization process of the second group, and so on. Figure 6.3 shows the Q_{MOS}×PESQ and the SRMR×PESQ plots for each group. The red circle in the Group 1 and Group 2 represents the point chosen as

Table 6.2: Grouping of the parameters and number of combinations for the enhancement process of algorithm A2.

Group	Parameters	Number of combinations
1	N_{FFT}, N_{win}	6
2	$n_e, G_{min,z_l}, G_{min,\nu}, \eta_{z_l}, \eta_\nu$	5808
3	$\beta^{dB}, \xi_{min,z_l}, \xi_{min,\nu}, \eta_{thresh}, \mu_\nu$	2880

preferred for these two groups. The combination of the three groups (entire feasible space for this algorithm) is depicted in Figure 6.4, in which the Pareto-optimal solutions are also shown.

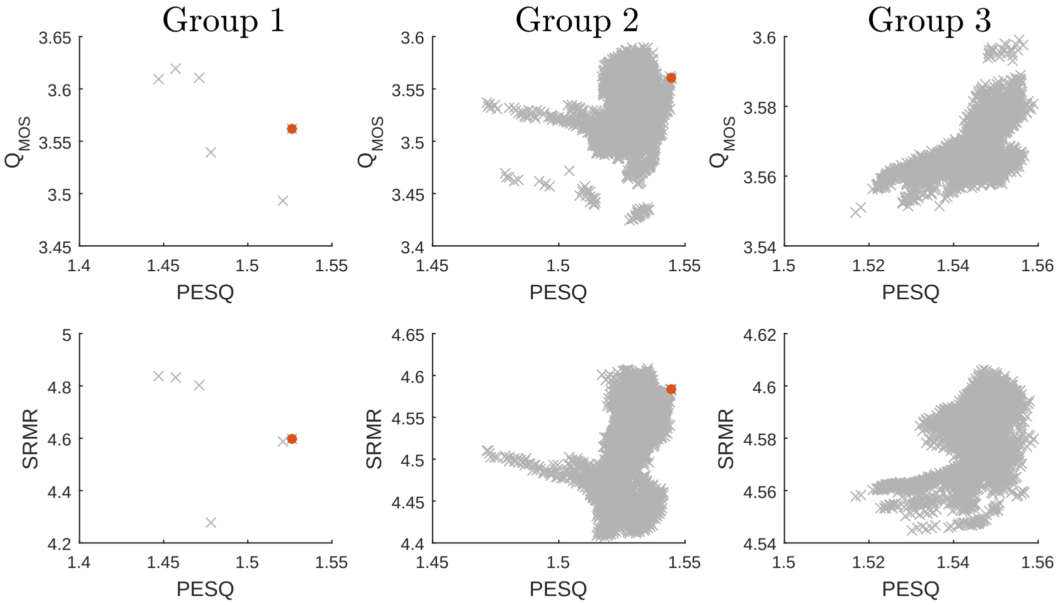


Figure 6.3: Q_{MOS}×PESQ and SRMR×PESQ for each group of parameters for the optimization process of algorithm A2. The red circle represents the preferred solution for groups 1 and 2.

6.2.2 Choice of the Optimal Operating Point

Figure 6.5 shows the operating points of the three original configurations of algorithm A2 in the same plot as the feasible objective space. As can be seen, the configuration that uses the STFT with a short window (A2.1) is already a Pareto-optimal solution. Unlike algorithm A1, it can be observed that the Q_{MOS} and SRMR measures are not equivalent in this case. Besides, it can be seen that the use of the STFChT (A2.3 scenario) improves the PESQ but reduces the Q_{MOS} and SRMR measures in comparison to the short-window STFT. The long-window STFT presents a little increase on the PESQ than the short-window, but this configuration also decrements the reverberant-based measures. This fact was already study

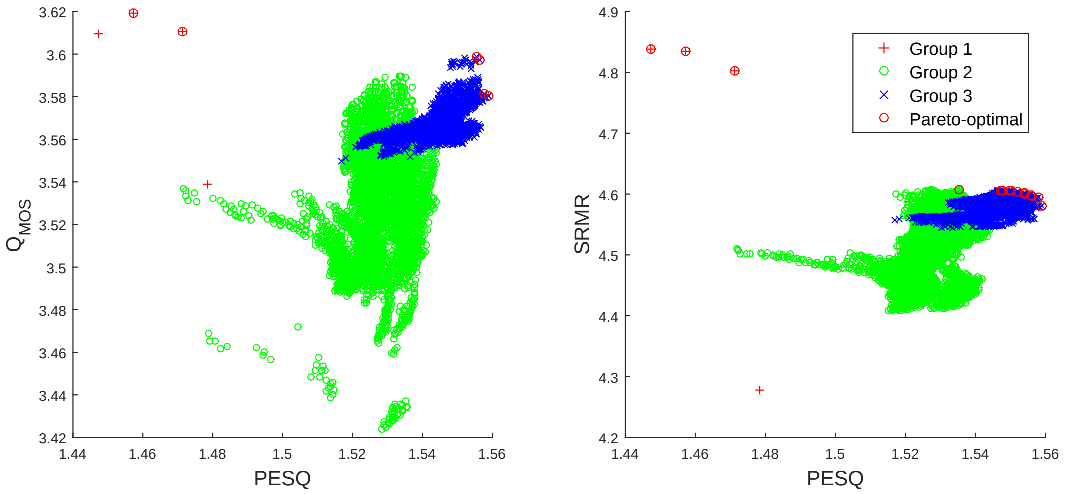


Figure 6.4: Feasible objective spaces of the optimization process of algorithm A2, using Q_{MOS}×PESQ and SRMR×PESQ measures.

in [21], where it was determined that processing in the STFChT domain results in less dereverberation at the output, but the enhanced speech does not suffer from addition of artifacts, as occurs when using the STFT domain. In either case, a new operating point for the algorithm among the Pareto-optimal solutions that lies in the region of high PESQ values can be chosen. Figure 6.6 zooms this region in the Q_{MOS}×PESQ and SRMR×PESQ plots. Points $C1$, $C2$, $C3$, $C4$, and $C5$ were chosen as candidates for optimal operating points of the algorithm.

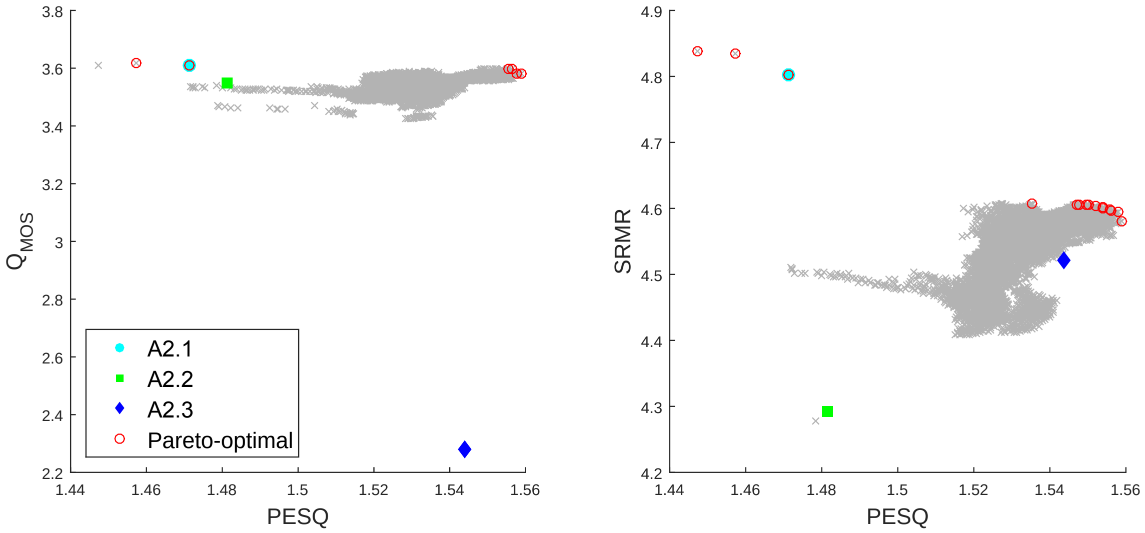


Figure 6.5: Original configurations of algorithm A2 represented in the Q_{MOS}×PESQ and SRMR×PESQ plots.

In order to determine the optimal operating point of algorithm A2, the WER of the configurations corresponding to points $C1$, $C2$, $C3$, $C4$, and $C5$ was calculated,

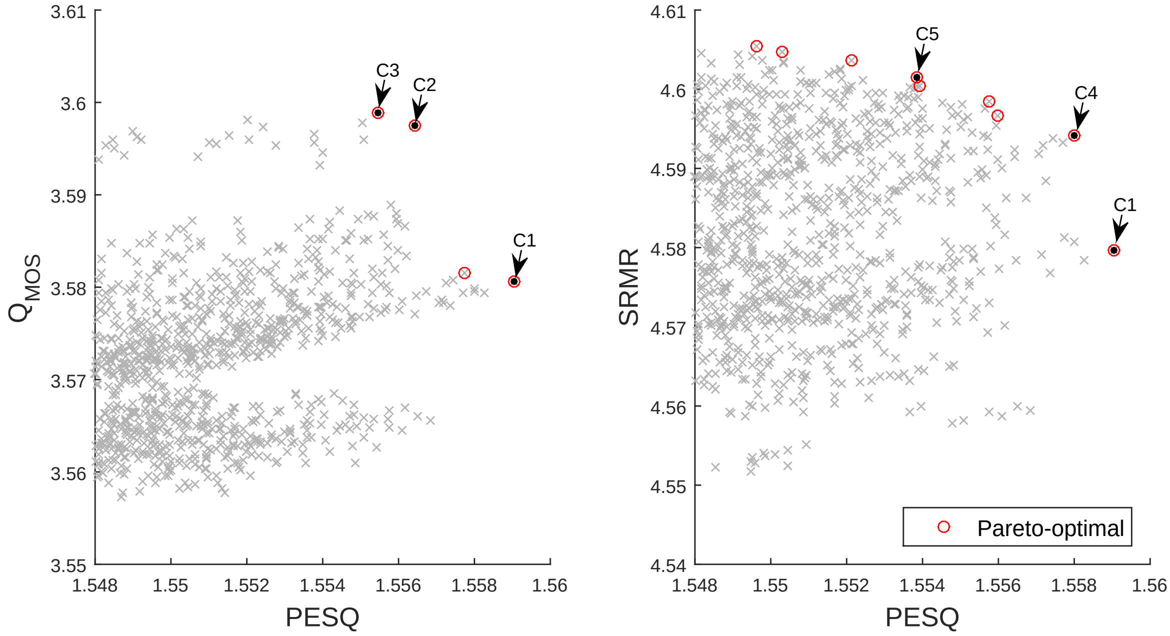


Figure 6.6: Candidates for optimal operating points of algorithm A2.

using the whole development dataset. Table 6.3 shows the average WER for these five candidates. Based on the results, point $C2$ is chosen as the optimal operating point for the algorithm, since it presents the lowest WER. A comparison between parameter values for the original and the optimal configurations of algorithm A2 is shown in Table 6.4. As can be observed, among all parameters, only seven of them change its value from the original to the optimized value, which means that the original algorithm was fairly well set.

Table 6.3: WER average in % corresponding to the operating points $C1$, $C2$, $C3$, $C4$, and $C5$ using the development dataset.

Operating Point	WER (%)
$C1$	47.65
$C2$	47.54
$C3$	47.94
$C4$	48.00
$C5$	47.71

6.3 Overall Results

The spectral subtraction based algorithm A2 was enhanced following the proposal of this work. Twelve parameters of this algorithm were finely tuned in order to optimize the Q_{MOS}, PESQ and SRMR measures simultaneously. Figure 6.7 shows the

Table 6.4: Comparison between the parameter values for the original and the optimal configurations for algorithm A2.

Parameter	Original value	Optimal value
N_{FFT}	–	1024
N_{win}	–	1024
n_e	768	768
G_{min,z_l}	0	0
$G_{min,\nu}$	-12	-5
β^{dB}	3	4
η_{z_l}	0.95	0.95
η_ν	0.95	0.95
ξ_{min,z_l}	-25	-20
$\xi_{min,\nu}$	-25	-15
η_{thresh}	0.3	0.4
μ_ν	0.98	0.89

Q_{MOS}×PESQ and SRMR×PESQ relations for the training set, marking the points of the original configurations (A2.1, A2.2 and A2.3), as well as the unprocessed and optimized points.

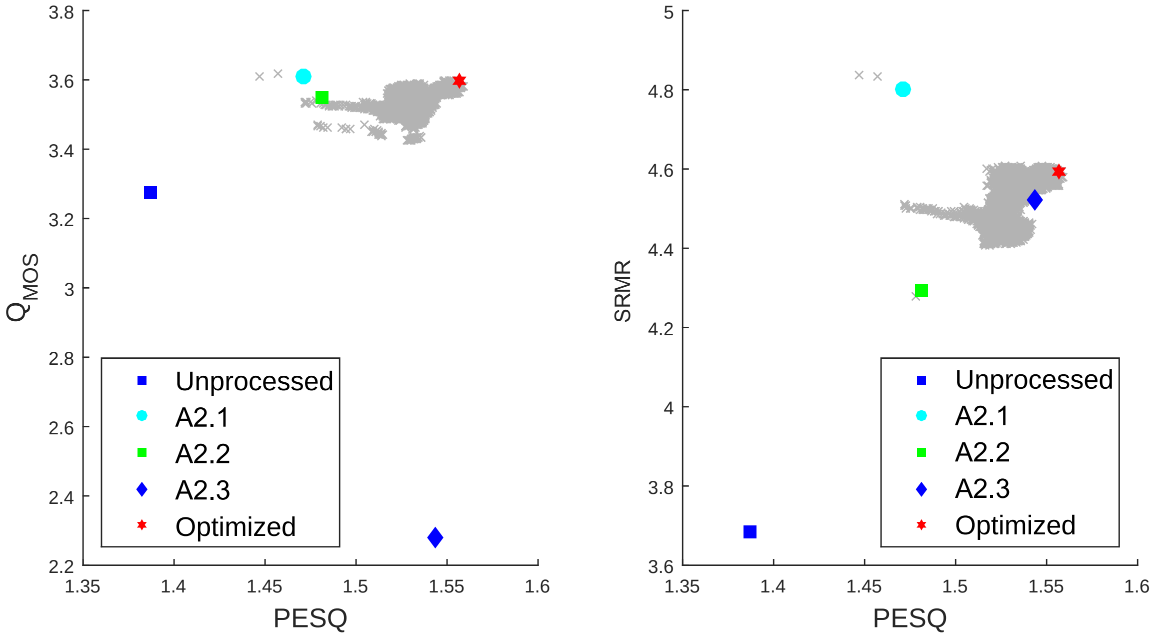


Figure 6.7: Q_{MOS}×PESQ (left) and SRMR×PESQ (right) plots for the training database, showing the unprocessed, original, and optimized operating points of algorithm A2.

For both development and evaluation datasets, Tables 6.5 and 6.6 compare, respectively, the PESQ, Q_{MOS}, SRMR, and WER measures of the unprocessed reverberant signals (Unprocessed scenario), the signals processed by the original configurations of algorithm A2 [21] (A2.1, A2.2 and A2.3 scenarios), and the signals processed by the enhanced algorithm (Optimized scenario). It can be appreciated that the training set generalizes well for the entire database. As can be seen from both tables, the optimized scenario presents the best WER in almost all reverberation conditions. In average, comparing with the original configurations A2.1, A2.2 and A2.3, the improvement in WER is about 10%, 3% and 8%, respectively. The fact that only half of the parameters changed their values from the original to the optimized configuration may explain why the improvement of WER is not very high.

It is important to notice that the optimized configuration rarely presents the best scores of SRMR, PESQ and Q_{MOS} values individually. The secret of the improvement in WER lies precisely in the joint optimization of these measures. For example, the STChT scenario presents the best PESQ scores in average; however, due to to the low value of Q_{MOS} that this configuration presents, the WER is not improved. Compared to algorithm A1, all measure values are lower for algorithm A2, which explains why the WER is higher when using this algorithm.

Table 6.5: Results of algorithm A2 for SimData and RealData development datasets. Bold numbers indicate the best results.

		SimData							RealData		
Measure	Scenario	Room 1		Room 2		Room 3		Avg.	Room 4		Avg.
		Near	Far	Near	Far	Near	Far	–	Near	Far	–
PESQ	Unprocessed	2.09	1.35	1.39	1.16	1.36	1.16	1.42	-	-	-
	A2.1	2.14	1.35	1.49	1.22	1.49	1.21	1.48	-	-	-
	A2.2	2.22	1.43	1.53	1.22	1.50	1.22	1.52	-	-	-
	A2.3	**2.43**	**1.47**	**1.65**	**1.23**	**1.60**	1.22	**1.60**	-	-	-
	Optimized	2.34	1.44	1.62	1.23	1.57	**1.22**	1.57	-	-	-
Q_{MOS}	Unprocessed	4.23	3.90	3.52	2.33	3.27	2.38	3.27	2.46	2.41	2.43
	A2.1	**4.28**	4.00	**3.85**	2.61	**3.53**	2.65	**3.49**	**3.77**	**3.71**	**3.74**
	A2.2	4.25	3.97	3.70	**2.66**	3.44	**2.75**	3.46	2.95	2.89	2.92
	A2.3	3.95	3.64	3.12	2.12	2.82	2.10	2.96	3.56	3.50	3.53
	Optimized	4.27	**4.02**	3.77	2.62	3.48	2.69	3.47	3.36	3.31	3.34
SRMR	Unprocessed	4.37	4.63	3.67	2.94	3.66	2.76	3.67	4.06	3.52	3.79
	A2.1	4.98	**5.66**	**4.72**	**4.48**	**4.78**	**4.39**	**4.83**	**7.76**	**6.85**	**7.30**
	A2.2	4.67	5.09	4.18	3.79	4.22	3.68	4.27	5.42	4.72	5.07
	A2.3	**4.99**	5.36	4.51	3.86	4.54	3.68	4.49	6.36	5.65	6.00
	Optimized	4.85	5.40	4.45	4.19	4.52	4.09	4.58	6.70	5.95	6.32
WER (%)	Unprocessed	**15.29**	25.29	43.90	85.80	51.95	88.9	51.81	88.71	88.31	88.51
	A2.1	26.35	40.41	40.52	71.26	48.05	75.74	50.36	78.98	77.38	78.18
	A2.2	15.56	**22.94**	33.15	73.8	40.18	75.27	43.45	76.48	75.73	76.10
	A2.3	18.02	27.24	35.99	79.69	41.57	83.18	47.58	80.85	80.79	80.82
	Optimized	18.12	25.54	**27.75**	**66.33**	**34.94**	**68.60**	**40.18**	**69.49**	**69.79**	**69.63**

Figures 6.8, 6.9, and 6.10 show comparative bar graphs for WER, Q_{MOS}, SRMR, and PESQ for all databases. The graphs of SRMR, Q_{MOS}, and PESQ also include the confidence intervals. These figures may offer a better perspective of the results obtained. For example, Figure 6.8 shows how the WER of the optimized scenario is improved in relation to the original configurations in all rooms except for Room 1. Figures 5.16 and 5.17 shows that all configurations of algorithm A2 have similar

Table 6.6: Results of algorithm A2 for SimData and RealData development datasets. Bold numbers indicate the best results.

		SimData							RealData		
Measure	Scenario	Room 1		Room 2		Room 3		Avg.	Room 4		Avg.
		Near	Far	Near	Far	Near	Far	–	Near	Far	–
PESQ	Unprocessed	2.14	1.60	1.40	1.19	1.37	1.17	1.48	-	-	-
	A2.1	2.23	1.65	1.55	1.27	1.53	1.23	1.57	-	-	-
	A2.2	2.25	1.70	1.57	1.26	1.52	1.23	1.59	-	-	-
	A2.3	**2.50**	**1.81**	**1.70**	**1.28**	**1.60**	1.24	**1.69**	-	-	-
	Optimized	2.40	1.75	1.67	1.28	1.58	**1.24**	1.65	-	-	-
Q_{MOS}	Unprocessed	4.24	3.96	3.61	2.38	3.20	2.40	3.30	2.51	2.57	2.54
	A2.1	4.27	4.05	**3.90**	2.66	**3.46**	2.65	3.50	**3.74**	**3.76**	**3.75**
	A2.2	4.27	4.04	3.81	**2.71**	3.42	**2.77**	3.50	2.99	3.03	3.01
	A2.3	3.98	3.69	3.23	2.20	2.71	2.10	2.98	3.48	3.47	3.47
	Optimized	**4.29**	**4.06**	3.86	2.69	3.41	2.71	**3.50**	3.40	3.41	3.41
SRMR	Unprocessed	4.50	4.58	3.74	2.97	3.57	2.73	3.68	3.17	3.19	3.18
	A2.1	5.13	**5.52**	**4.89**	**4.52**	**4.82**	**4.37**	**4.87**	**6.06**	**6.01**	**6.04**
	A2.2	4.81	5.01	4.31	3.81	4.23	3.62	4.30	4.17	4.25	4.21
	A2.3	**5.19**	5.30	4.70	3.93	4.56	3.63	4.55	4.88	4.84	4.86
	Optimized	4.99	5.32	4.61	4.22	4.55	4.04	4.62	5.17	5.29	5.23
WER (%)	Unprocessed	**18.06**	25.38	42.98	82.20	53.54	88.04	51.68	89.72	87.34	88.53
	A2.1	33.09	43.09	41.36	64.62	51.60	77.94	51.93	82.85	79.17	81.01
	A2.2	18.21	**24.28**	32.26	64.54	41.19	73.72	42.35	81.44	76.37	78.90
	A2.3	20.65	27.65	35.20	72.47	45.09	81.62	47.09	84.25	80.25	82.25
	Optimized	21.25	27.77	**27.75**	**57.44**	**38.01**	**68.50**	**40.10**	**76.88**	**73.80**	**75.34**

scores in the considered measures. The optimized configuration does not outstand in any particular measure, but presents the best compromise between them.

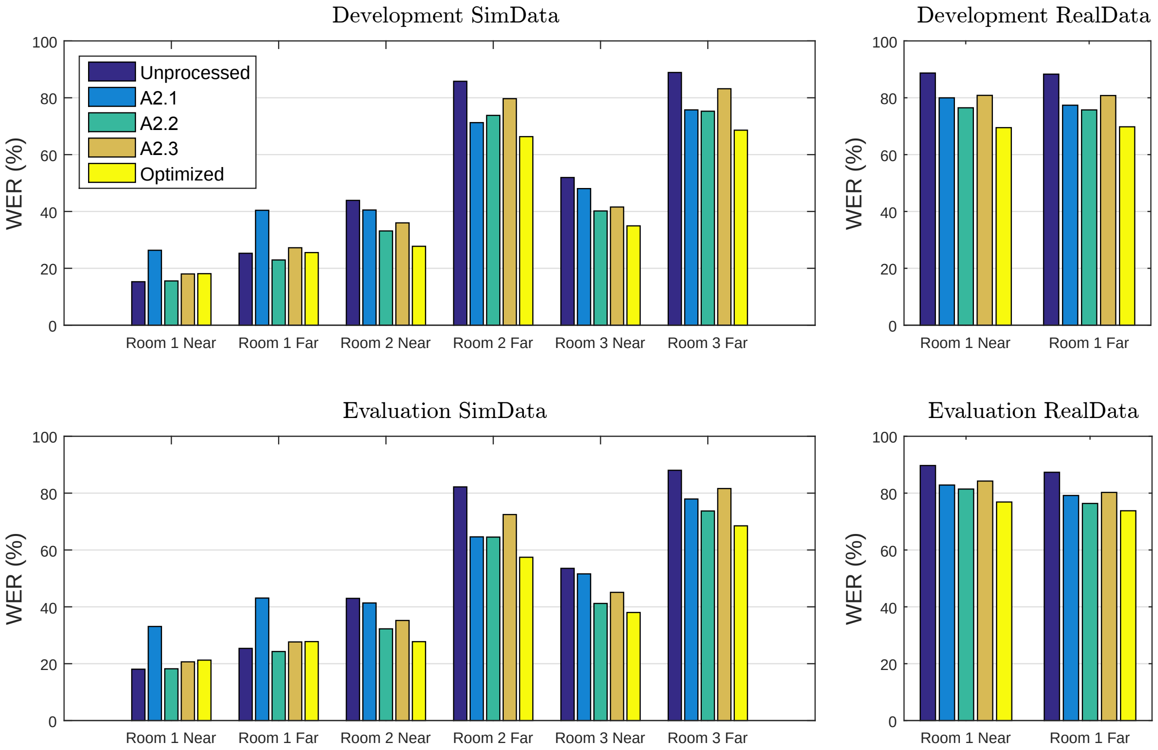


Figure 6.8: Word error rate (WER) results for the whole database comparing the different configurations of algorithm A2.

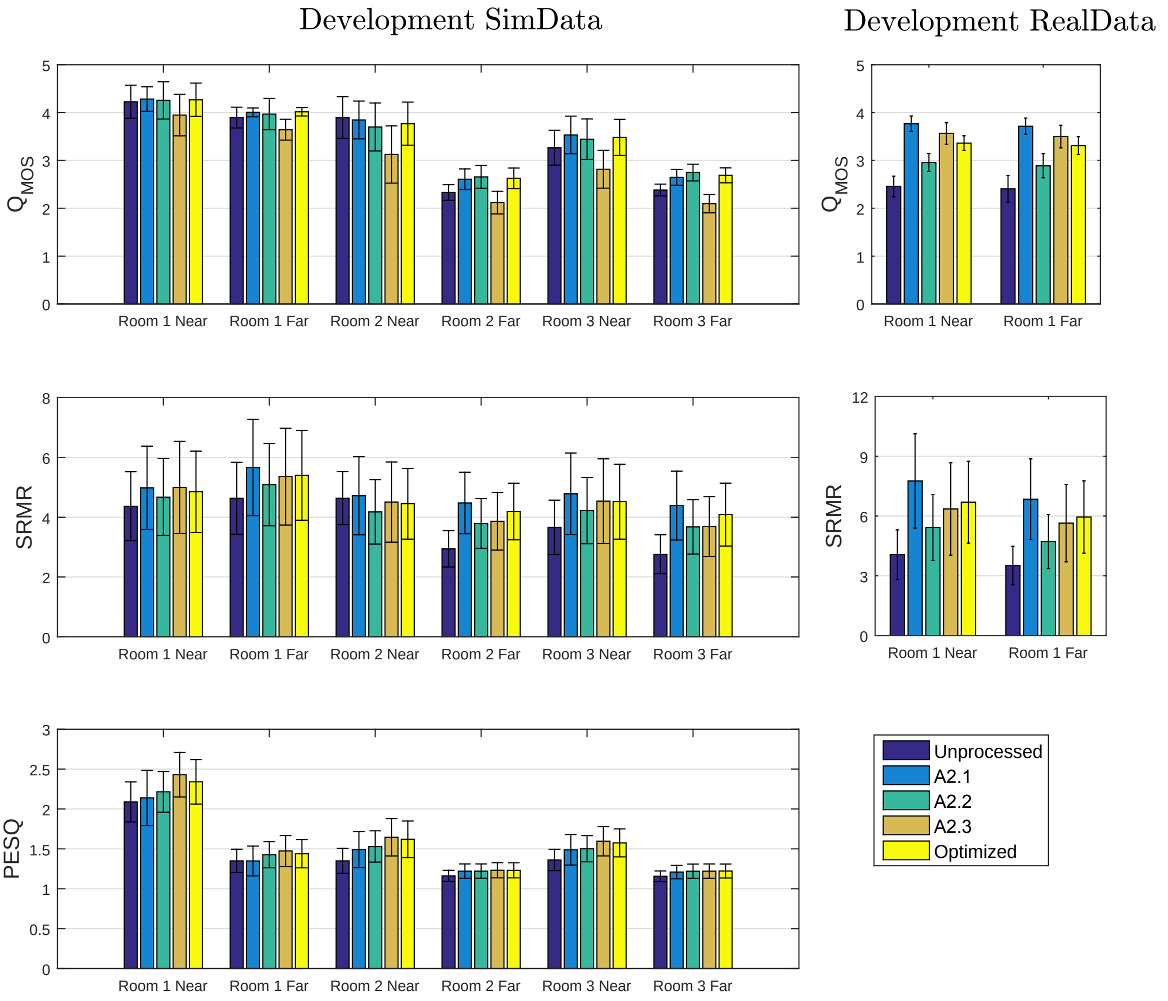
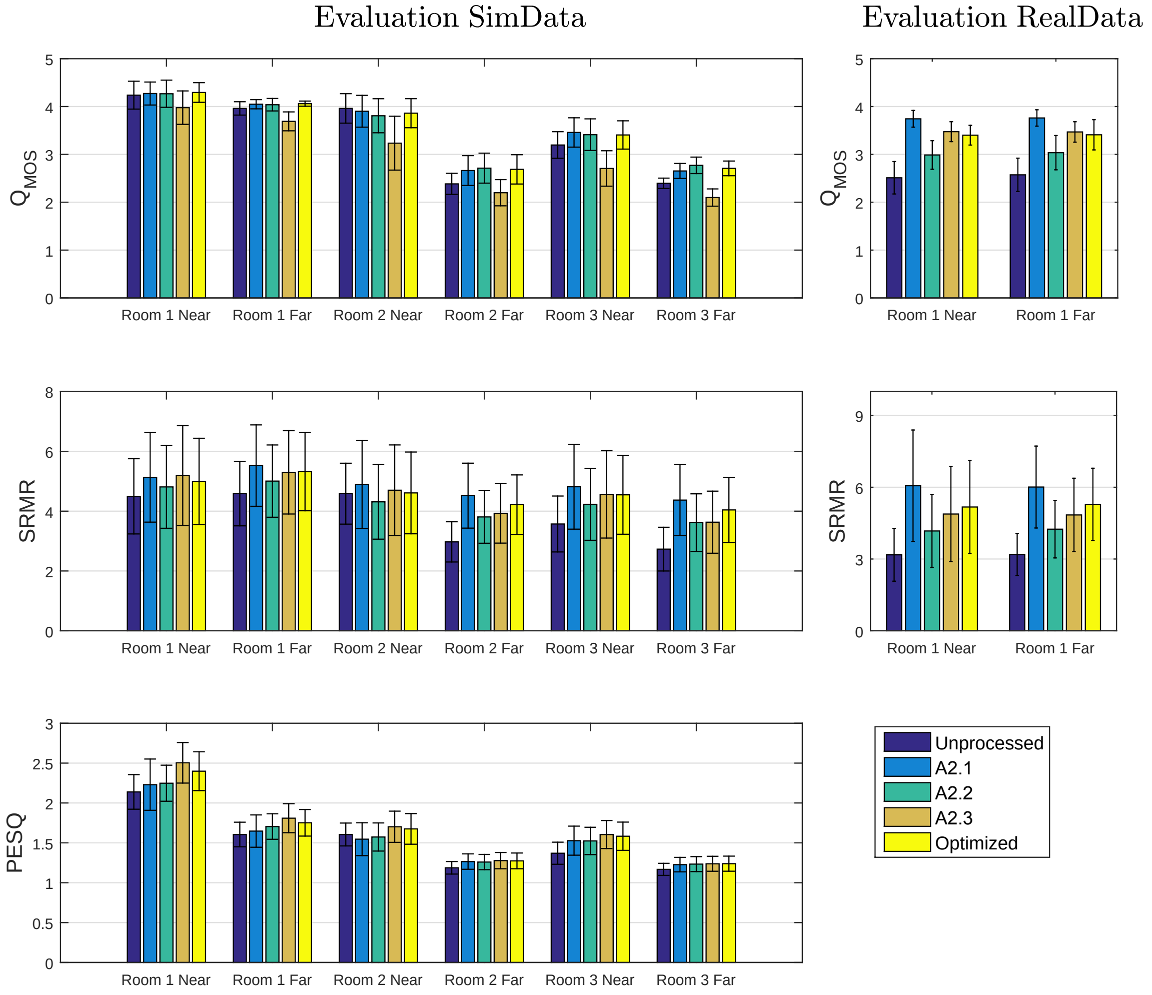


Figure 6.9: Q_{MOS}, SRMR and PESQ results for SimData and RealData development dataset comparing the different configurations of algorithm A2.


Evaluation SimData

Evaluation RealData

Figure 6.10: Q_{MOS}, SRMR and PESQ results for SimData and RealData development dataset comparing the different configurations of algorithm A2.

6.4 Conclusion

The dereverberation and noise reduction algorithm A2 was enhanced in this chapter following the proposed methodology. This enhancement was implemented by combining the Q_{MOS}, SRMR and PESQ measures as objectives to be maximized. Twelve of the algorithm parameters were optimized and the enhanced setup was compared with the three original configurations of the algorithm. Table 6.7 compares the original and enhanced configurations of this algorithm in terms of parameter values and total average WER achieved by the ASR system. Results demonstrate effectiveness as the optimized scenario outperformed the other configurations in terms of speech intelligibility, as assessed by the lower WER achieved by the ASR system. The difference of WER between the original setups of the algorithm and the enhanced scenario is not very high, since the original setup of the parameters was well established.

Table 6.7: Comparison of the parameter values and the total average WER between the original and the enhanced configurations of algorithm A2.

	Original Configuration	Enhanced Configuration
Parameters	$N_{FFT} = \{512, 3262\}$ $N_{win} = \{512, 2048\}$ $n_e = 768$ $G_{min,z_l} = 0$ $G_{min,\nu} = -12$ $\beta^{dB} = 3$ $\eta_{z_l} = 0.95$ $\eta_\nu = 0.95$ $\xi_{min,z_l} = -25$ $\xi_{min,\nu} = -25$ $\eta_{thresh} = 0.3$ $\mu_\nu = 0.98$	$N_{FFT} = 1024$ $N_{win} = 1024$ $n_e = 768$ $G_{min,z_l} = 0$ $G_{min,\nu} = -5$ $\beta^{dB} = 4$ $\eta_{z_l} = 0.95$ $\eta_\nu = 0.95$ $\xi_{min,z_l} = -20$ $\xi_{min,\nu} = -15$ $\eta_{thresh} = 0.4$ $\mu_\nu = 0.89$
WER(%)	A2.1: 55.06 A2.2: 47.40 A2.3: 51.80	44.50

Chapter 7

Enhancement of Dereverberation Algorithm A3

In this chapter the third and last dereverberation algorithm is enhanced following the proposed method. This algorithm, which will be denoted as algorithm A3, was introduced in [72] and proposes a single-channel speech enhancement method using zero phase transformation, which is defined as the inverse DFT of a spectral amplitude. This chapter follows the same structure as the two previous chapters, that is, algorithm A3 is presented first, introducing the parameters involved in the optimization process, which is detailed next. Finally, comparative results between the original and enhanced configuration are shown.

7.1 A3 Description

Algorithm A3 aims at the reduction of noise and reverberation using the zero phase transform. The zero phase version $y_{zp}(n)$ of a speech signal $y(n)$ is computed as

$$y_{zp}(n) = \mathrm{IDFT}(|Y(e^{j\omega})|^{\beta}), \tag{7.1}$$

where $|Y(e^{j\omega})|$ is the magnitude of the Fourier transform of $y(n)$, β is an integer, and IDFT stands for the inverse DFT. Since the spectral amplitude of a reverberant and noisy sequence is approximately flat, its zero phase signal takes nonzero values only around the origin. This behavior allows to detect the reverberation location and remove it. Thus, algorithm A3 computes the zero phase version of the reverberant and noisy signal and then replaces the reverberant samples.

A diagram of algorithm A3 is depicted in Figure 7.1. First, the reverberant and noisy speech signal is filtered with a pre-emphasis filter, and then transformed into the STFT domain using a Hamming window of 32 ms with an overlap of 10 ms. The zero phase version $y_{zp}(n)$ is computed with $\beta = 1$. The revereberant samples substitution is then computed for each frame of the $y_{zp}(n)$ sequence, using the knowledge that the reverberation is located in the first period of the zero phase sequence. Thus the second period is used for replacing the corrupted part of the sequence. A peak selection algorithm is used here for obtaining the period of the voiced speech. The amount of reverberant samples to be replaced is determined empirically as $L = 10$ samples. The speech sequence is reconstructed using the new magnitude with the original phase, and a overlap-add algorithm is finally performed in order to convert the separated frames into a temporal sequence again.

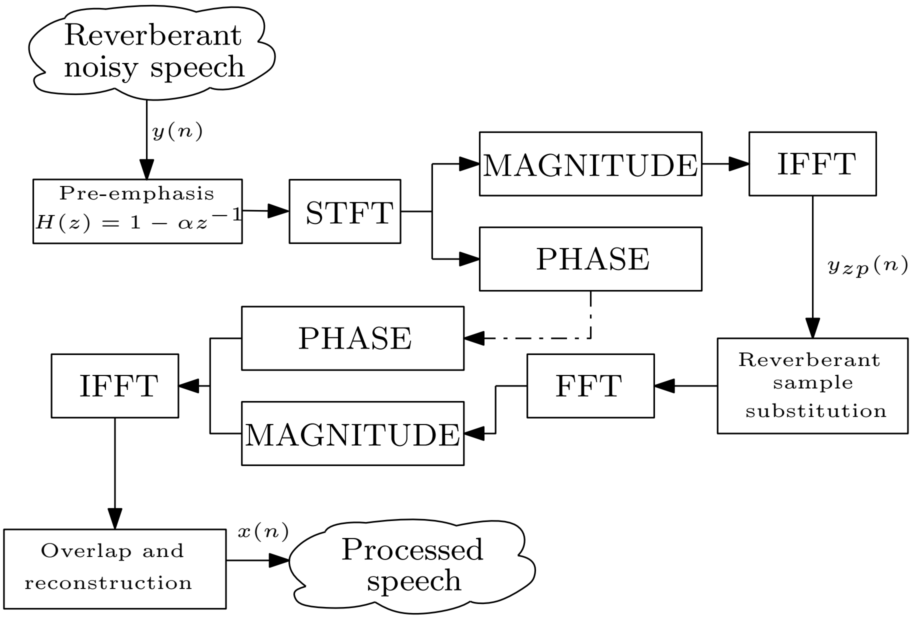


Figure 7.1: Diagram of algorithm A3.

7.2 Enhancement Process

As in the two previous chapters, this section describes the methodology used for the enhancement process of algorithm A3. First, the selected parameters and their range of analysis are described. Then, a new operating point for the algorithm is chosen so that the set of selected measures are optimized.

7.2.1 Choice of the Parameters

For algorithm A3, five parameters were chosen for the optimization process. These parameters, along with their description, original values and range of analysis are detailed in Table 7.1. The combination of these parameters within their range gives a total of 9240 different setups for algorithm A3.

Table 7.1: Parameters chosen for the optimization process of algorithm A3, including a brief description, their original values, and the range of analysis.

Parameter	Description	Original value	Range
α	Position of the zero of the pre-emphasis filter.	0.7	$\{0, 0.1, 0.2, 0.3, \ldots, 0.9, 0.9375\}$
β	Integer used to compute the zero phase sequence.	1	$\{1, 2\}$
L	Amount of reverberant samples to be replaced.	10	$\{5, 6, 7, 8, 9, 10, \ldots, 25\}$
t_L	Start of the interval for the pitch detection in ms.	2	$\{1, 2, 3, 4\}$
t_H	End of the interval for the pitch detection in ms.	8	$\{6, 7, 8, 9, 10\}$

7.2.2 Choice of the Optimal Operating Point

Figure 7.2 shows the feasible objective spaces on the Q_{MOS}×PESQ and the SRMR×PESQ plots. Based on this figure, one notices that algorithm A3 may not be able to improve the quality and intelligibility of the considered speech signals, since the PESQ values for all setups is considerably worse than the PESQ value of the unprocessed signals. As seen in Chapter 5, an improvement on reverberant-based measures is only significant when the values of PESQ are also improved. It seems clear that algorithm A3 introduces artifacts that degrades the speech signals. This fact can be also appreciated on the Q_{MOS}×PESQ plot, in which the values of Q_{MOS} for the processed are far away from the $1 - 5$ MOS scale. This is because the version of the QAreverb measure used here was adapted to deal with non-distorted dereverberated signals. Figure 7.3 shows the function employed for mapping the measure Q into the Q_{MOS}, as described in Equation 3.3. As can be seen, the range of Q that correctly maps this measure into the MOS scale is -26.5 to 0.6. Outside this range, the values of Q_{MOS} can be too distant from the MOS scale. Therefore, the use of the Q_{MOS} measure will be unconsidered for algorithm A3.

Five candidates for the new operating point of algorithm A3 were chosen among the Pareto-optimal solutions in the SRMR×PESQ feasible objective region. Figure 7.4 show these points, labeled as $\{D1, D2, D3, D4, D5\}$.

In order to determine the optimal operating point of algorithm A3, the WER of the configurations corresponding to points $D1$, $D2$, $D3$, $D4$, and $D5$ was calculated, using the whole development dataset. Table 7.2 shows the average WER for these five candidates. Based on the results, point $D2$ is chosen as the optimal operating point for the algorithm, since it presents the lowest WER. A comparison between the parameter values for the original and the optimal configurations for algorithm A3 is shown in Table 7.3.

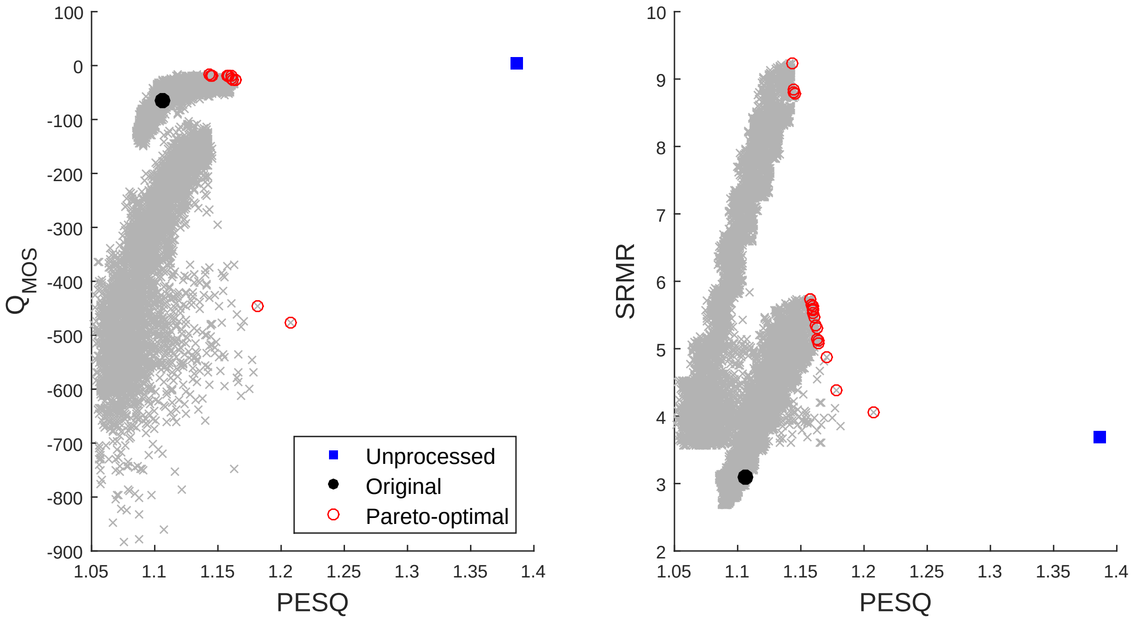


Figure 7.2: Q_{MOS}×PESQ and SRMR×PESQ for the optimization process of algorithm A3.

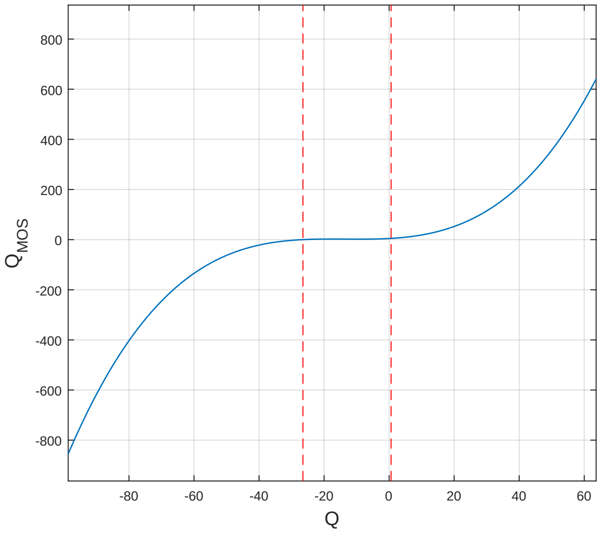


Figure 7.3: Function that maps the Q measure into the Q_{MOS}.

7.3 Overall Results

The dereverberation algorithm A3 was enhanced following the proposal of this work. Five parameters of this algorithm were finely tuned in order to optimize the PESQ and SRMR measures simultaneously. Figure 7.5 shows the SRMR×PESQ relations for the training set, marking the points of the original configuration, as well as the unprocessed and optimized points. As can be seen, although the SRMR value for

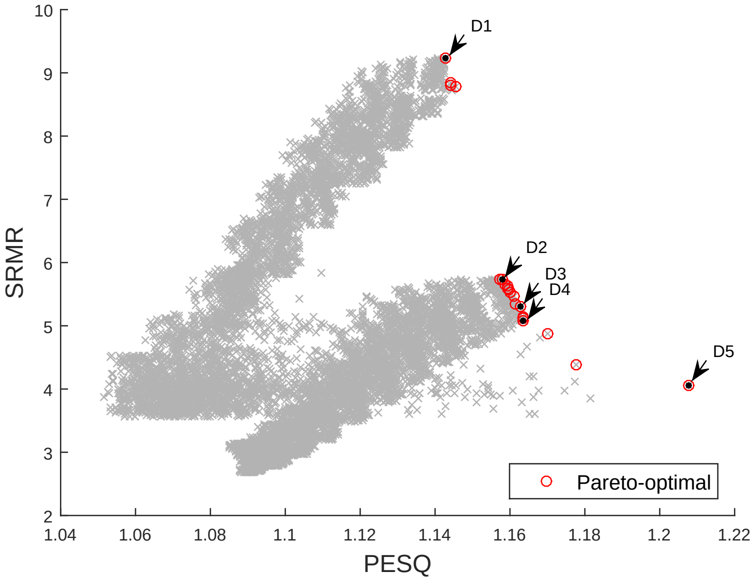


Figure 7.4: Candidates for optimal operating point of algorithm A3.

Table 7.2: WER average in % corresponding to operating points $D1$, $D2$, $D3$, $D4$, and $D5$ using the development dataset.

Operating Point	WER (%)
$D1$	92.57
$D2$	81.05
$D3$	81.21
$D4$	81.15
$D5$	91.75

Table 7.3: Comparison between the parameter values for the original and the optimal configurations of algorithm A3.

Parameter	Original value	Optimal value
α	0.7	0
β	1	1
L	10	25
t_L	2	1
t_H	8	10

the optimized configuration is higher than the SRMR value of the original and unprocessed scenarios, the PESQ value decreases when compared with the unprocessed scenario. Thereby it is not expected an improvement in the intelligibility of speech signals processed by algorithm A3. This fact is confirmed after calculating the WER through the ASR system. These WER results, together with the PESQ and SRMR values for the development and evaluation datasets, are shown in Tables 7.4 and 7.5.

As can be observed from these tables, the WER achieved by the optimized configuration, although it improves in average the original configuration of algorithm A3, is worse than the unprocessed signals for every reverberant condition.

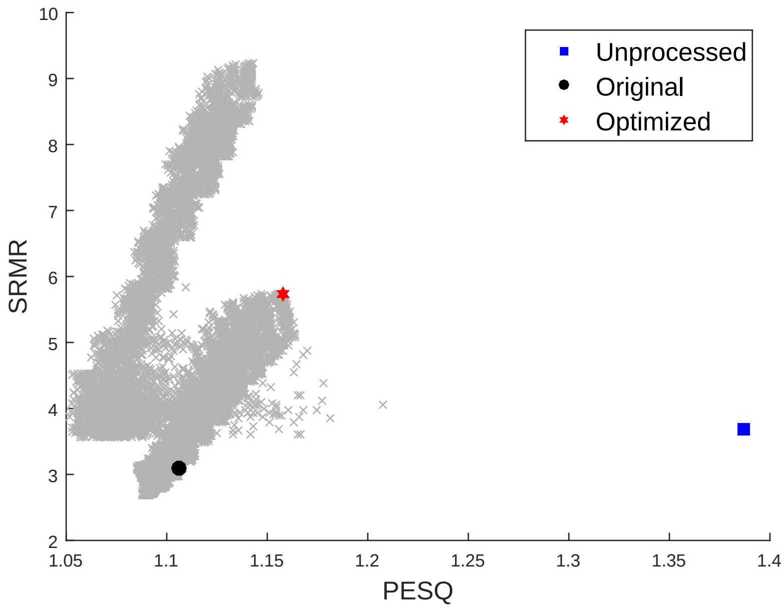


Figure 7.5: SRMR×PESQ plot for the training database, showing the unprocessed, original, and optimized operating points of algorithm A3.

Table 7.4: Results for SimData and RealData development datasets of algorithm A3. Bold numbers indicate the best results.

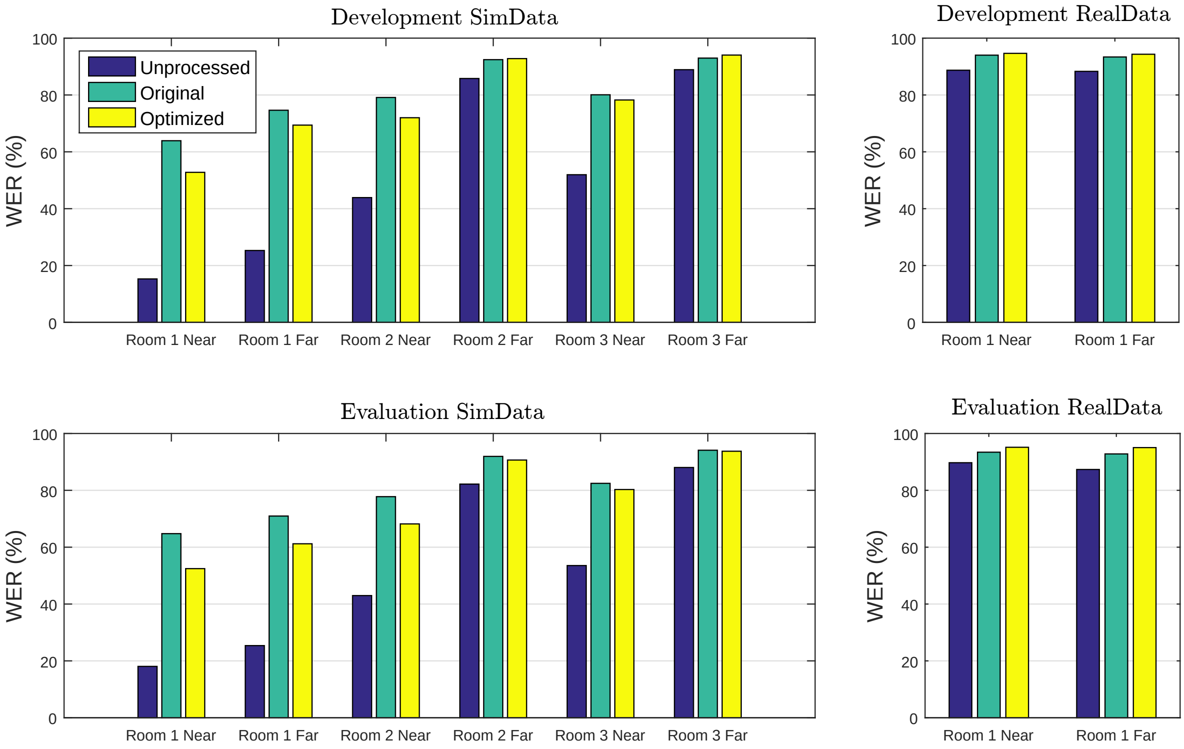
		SimData							RealData		
Measure	Scenario	Room 1		Room 2		Room 3		Avg.	Room 1		Avg.
		Near	Far	Near	Far	Near	Far	–	Near	Far	–
PESQ	Unproc.	**2.09**	**1.35**	**1.39**	**1.16**	**1.36**	**1.16**	**1.42**	-	-	-
	Original	1.17	1.12	1.13	1.09	1.13	1.08	1.12	-	-	-
	Opt.	1.28	1.16	1.20	1.13	1.20	1.12	1.18	-	-	-
SRMR	Unproc.	4.37	4.63	3.67	2.94	3.66	2.76	3.67	4.06	3.52	3.79
	Original	3.44	3.67	2.99	2.59	3.14	2.51	3.06	4.63	4.22	4.42
	Opt.	**6.85**	**7.34**	**6.23**	**4.79**	**5.90**	**4.57**	**5.94**	**6.42**	**5.51**	**5.96**
WER (%)	Unproc.	**15.29**	**25.29**	**43.90**	**85.80**	**51.95**	**88.9**	**51.81**	**88.71**	**88.31**	**88.51**
	Original	63.91	74.63	79.12	92.43	80.07	92.98	80.50	94.01	93.37	93.69
	Opt.	52.78	69.40	72.00	92.80	78.24	94.04	76.52	94.95	94.33	94.64

Figure 7.6 shows the same results as the above tables with bar graphs. As can be seen, the unprocessed scenario has the lowest WER in every reverberant condition. For the SimData set, the optimized configuration outperforms the original setup, presenting an average improvement of 4% for the development dataset and 6% for the evaluation dataset. For the RealData set, however, the original configuration presents a slightly better WER than the optimized configuration. It is important to notice that all the optimization process is carried out with the SimData set, which

Table 7.5: Results of algorithm A3 for SimData and RealData evaluation datasets. Bold numbers indicate the best results.

		SimData							RealData		
Measure	Scenario	Room 1		Room 2		Room 3		Avg.	Room 1		Avg.
		Near	Far	Near	Far	Near	Far	–	Near	Far	–
PESQ	Unproc.	**2.14**	**1.60**	**1.40**	**1.19**	**1.37**	**1.17**	**1.48**	-	-	-
	Original	1.17	1.16	1.13	1.09	1.13	1.09	1.13	-	-	-
	Opt.	1.30	1.26	1.21	1.14	1.21	1.13	1.21	-	-	-
SRMR	Unproc.	4.50	4.58	3.74	2.97	3.57	2.73	3.68	3.17	3.19	3.18
	Original	3.49	3.80	3.08	2.68	3.12	2.65	3.13	3.71	3.86	3.78
	Opt.	**7.11**	**6.86**	**6.47**	**4.84**	**5.85**	**4.44**	**5.93**	**5.36**	**5.17**	**5.26**
WER (%)	Unproc.	**18.06**	**25.38**	**42.98**	**82.80**	**53.54**	**88.04**	**51.68**	**89.72**	**87.34**	**88.53**
	Original	64.77	70.96	77.78	91.94	82.46	94.11	80.32	93.42	92.81	93.11
	Opt.	52.46	61.20	68.19	90.66	80.29	93.76	74.41	95.15	95.04	95.09

can explain this fact. Figures 7.7 and 7.8 show the comparative bar graphs for the SRMR and PESQ for the development and evaluation datasets, respectively. As can be seen, the SRMR values of the optimized configuration are higher than the SRMR values of the original configuration and the unprocessed scenario in every reverberant condition. The PESQ values of the optimized configuration are also higher than the ones for the original configuration, however they are lower than for the unprocessed scenario, which explain why the WER could not be improved for this particular algorithm.


Development SimData

Development RealData

Evaluation SimData

Evaluation RealData

Figure 7.6: Word error rate (WER) results for the whole database comparing the different configurations of algorithm A3.

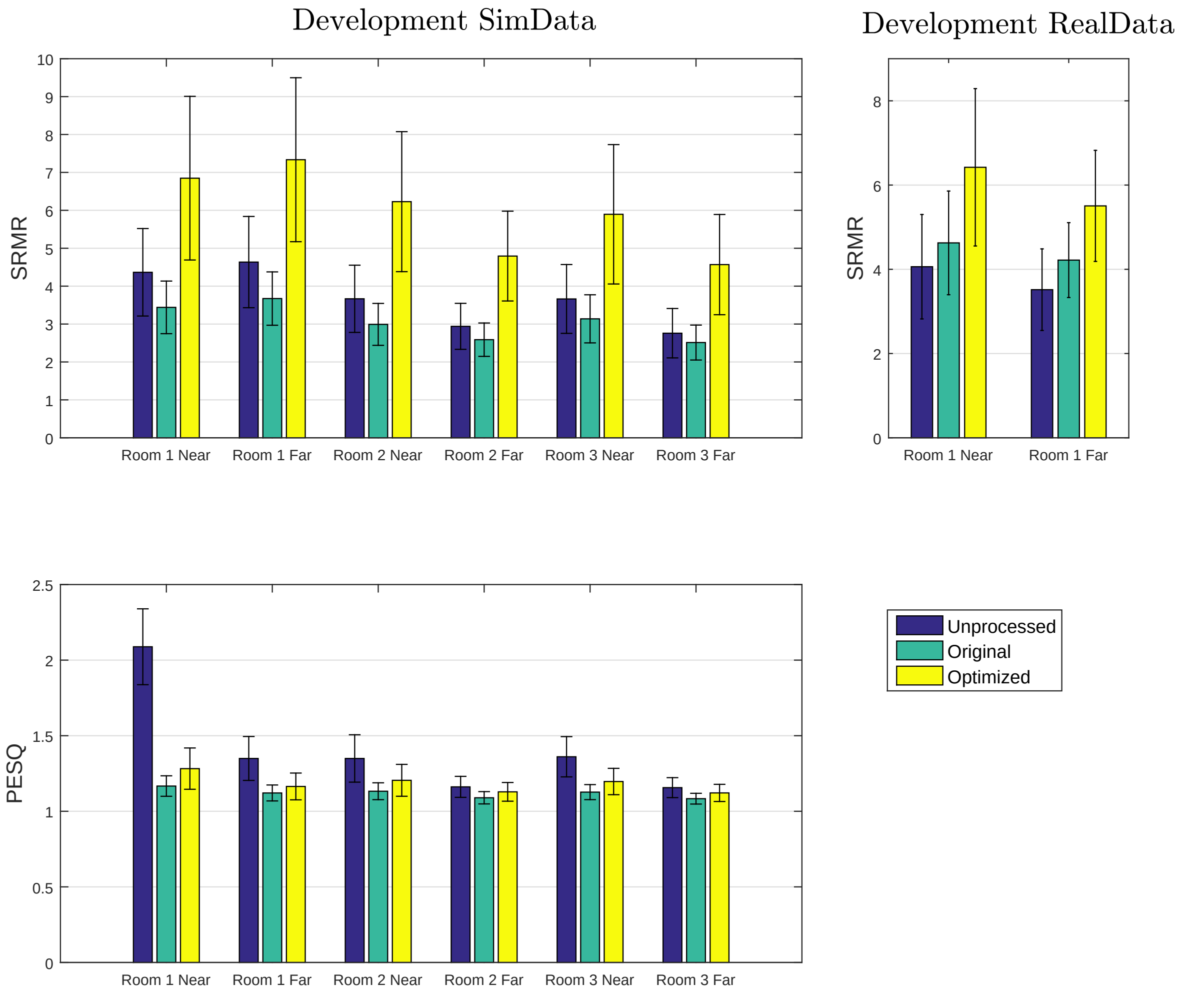


Figure 7.7: SRMR and PESQ results for SimData and RealData development dataset comparing the different configurations of algorithm A3.

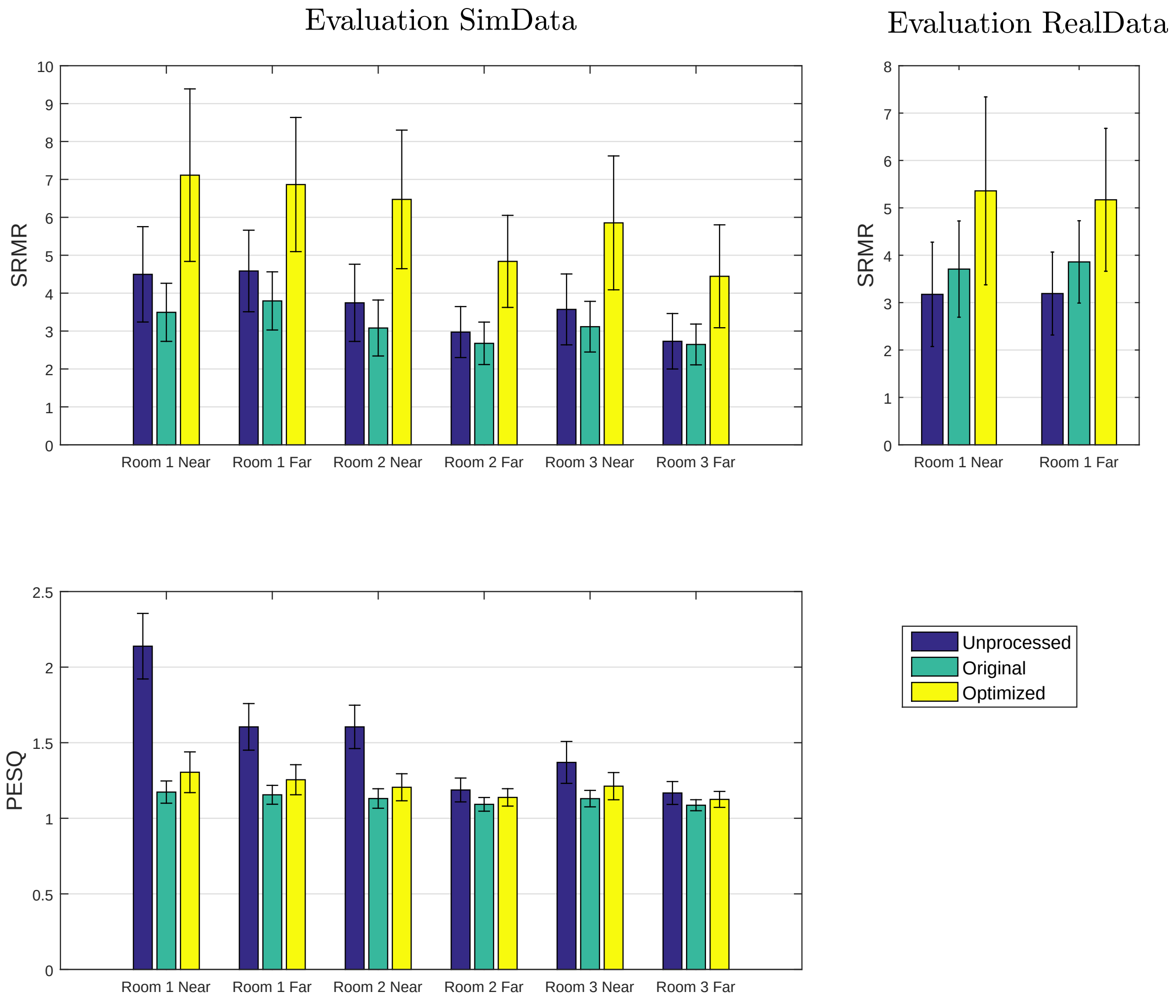


Figure 7.8: SRMR and PESQ results for SimData and RealData development dataset comparing the different configurations of algorithm A3.

7.4 Conclusion

In this chapter algorithm A3, which intends to reduce noise and reverberation through the zero transform, was enhanced following this work proposal. Table 7.6 compares the original and enhanced configuration of this algorithm in terms of parameter values and total average WER achieved by the ASR system. Although the WER achieved by the enhanced version of the algorithm outperformed the original configuration, this algorithm does not seem to be valid for dereverberation issues, since it introduces artifacts that degrade the processed speech signals. This fact could be observed even before computing the WER of the ASR system. By observing the feasible region in the multi-objective plot, it is possible to know the effectiveness of the algorithm. In this case, the lower PESQ achieved by the signals processed by algorithm A3 reveled the limited capacity of this algorithm to desreverberate speech signals.

Table 7.6: Comparison of the parameter values and the total average WER between the original and the enhanced configuration of algorithm A3.

	Original Configuration	Enhanced Configuration
Parameters	$\alpha = 0.7$	$\alpha = 0$
	$\beta = 1$	$\beta = 1$
	$L = 10$	$L = 25$
	$t_L = 2$	$t_L = 1$
	$t_H = 8$	$t_H = 10$
WER(%)	82.08	77.84

Chapter 8

Conclusions and Future Work

This work presented an enhancement procedure for dereverberation algorithms based on the simultaneous optimization of several assessment measures. The proposed method was applied to three different dereverberation algorithms, which had been proposed in the REVERB challenge [6]. The results achieved were promising and satisfactory, since the WER of every enhanced version of the three algorithms outperformed the original configurations. The proposed technique worked fine when using the SRMR or Q_{MOS} reverberant-measures combined with the PESQ (overall quality measure). Figure 8.1 shows different regions in the $Q_{MOS}\times$PESQ objective space in which the WER of some points were calculated. As can be seen, the region with lower WER corresponds to higher values of PESQ and Q_{MOS} simultaneously. A similar plot would be obtained by changing the Q_{MOS} measure for the SRMR. This figure also shows that, for lower values of PESQ, the increment of the reverberant-based measures worsens the WER drastically.

Although any other measures can be used with the proposed methodology, the three measures mentioned above have proved to be effective in practice. Table 8.1 compares the WER of the original configuration of the three algorithms with the WER of the enhanced algorithms. This WER was obtained as the average of the entire database. Algorithm A1 was significantly enhanced, with a 22% reduction in the WER. Algorithm A2 presented an improvement of 10.56%, 2.9% and 7.3% in relation over its three original configurations. Lastly, algorithm A3 presented a WER reduction of 4.24%. Based on these results it can be also concluded that the more effective dereverberation algorithm is algorithmA1.

It is important to mention that the main purpose of this work was to study and show a methodology to enhance dereverberation algorithms, not the enhancement itself. More effort could be made in order to further improve these algorithms, such as expanding the range of analysis of the parameters, or increasing the number of parameters. This could be suggested as future work and may require more computational power.

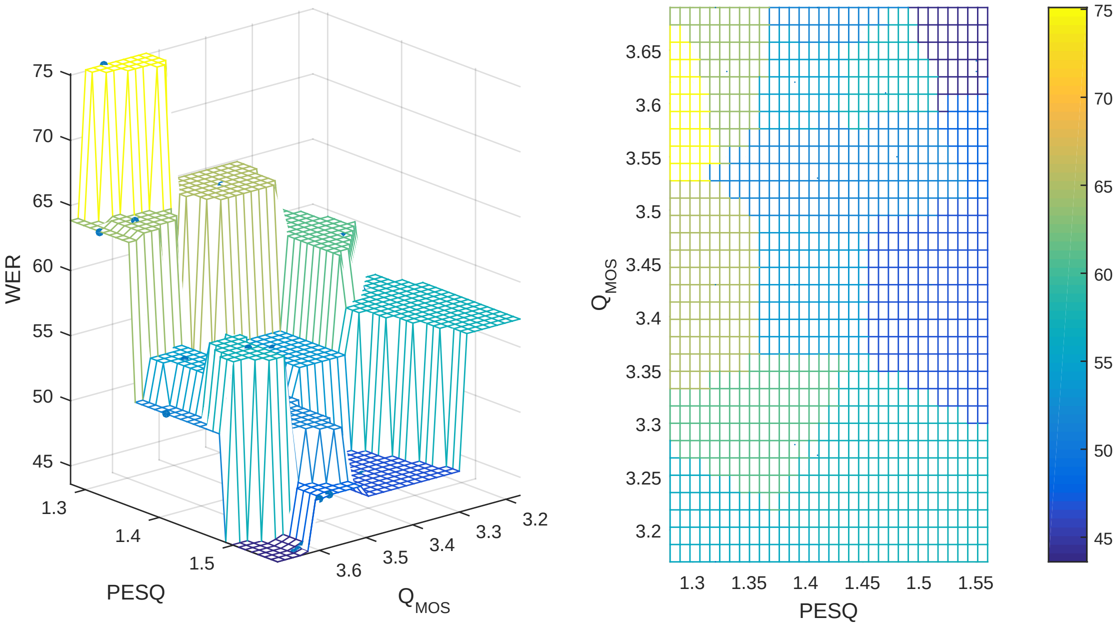


Figure 8.1: WER for the different regions of Q_{MOS} and PESQ.

Table 8.1: Comparison of the WER achieved by the original and enhanced versions of the algorithms A1, A2, and A3.

Algorithm	WER (Original)	WER (Enhanced)
A1	63.81%	41.59%
A2.1	55.06%	44.50%
A2.2	47.40%	
A2.3	51.80%	
A3	82.08%	77.84%

The proposed method can be also useful to compare the performance of different dereverberation algorithms. The location of the operating point of the algorithm in the objective feasible space gives an accurate knowledge about the algorithm's competency, and the extent to which it can be improved. By way of example, Figure 8.2 shows the optimal operating points of the algorithms A1, A2 and A3 in the SRMR×PESQ plot. The total average WER for each of these algorithms is also depicted in this figure. Besides, another point, labeled as A4, corresponding to a non-existent algorithm is shown. This representation allows one to compare and evaluate the performance of any dereverberation algorithm even without knowing the WER of processed signals. In this way, algorithm A4 is expected to outperform all other algorithms in terms of speech intelligibility.

Furthermore, many other measures can be employed in the optimization process. Figure 8.3 shows an example of a three-dimensional plot considering the SRMR,

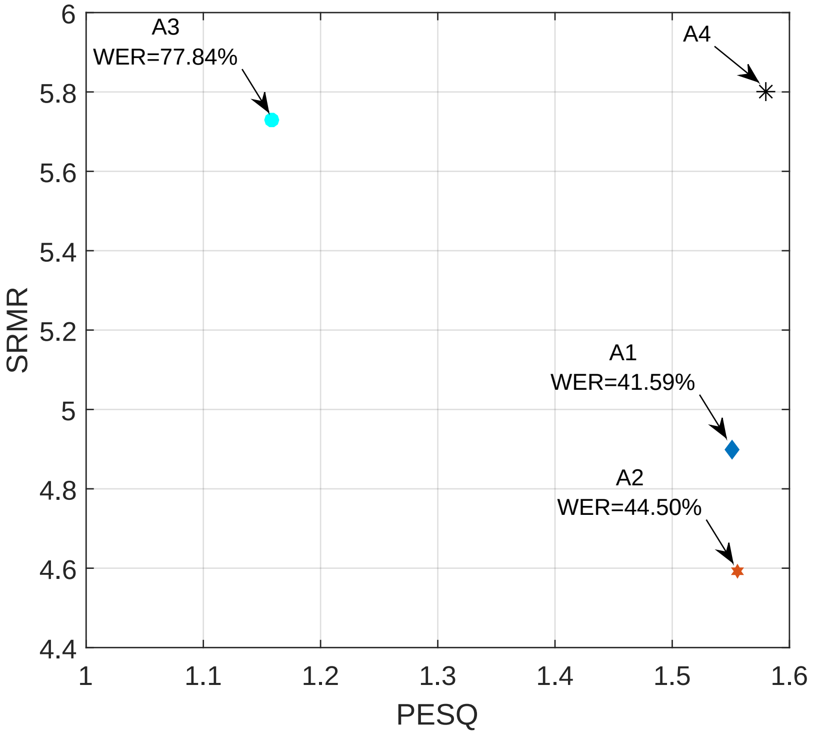


Figure 8.2: Comparison of the operating points of algorithms A1, A2 and A3 in the SRMR×PESQ plot.

Q_{MOS} and PESQ as objectives, using the feasible space of algorithmA1. It is important to notice that the more the number of measures used, the more the cardinality of the non-dominated solutions is. If the number of measures is greater than three, it will be more difficult to reduce the Pareto-optimal set in the absence of a graphical representation. However, it is worth a deeper study on the combination of different metrics.

Another possibility is to form a composite objective O as the weighted sum of different measures, that is,

$$O = \mathrm{w}_1 O_1 + \mathrm{w}_2 O_2 + \cdots \mathrm{w}_M O_M, \tag{8.1}$$

where the weights w_i, for $i = 1, \cdots, M$, are proportional to a preference factor assigned to each measure O_i. This method converts the multi-objective optimization problem into a single-objective optimization problem. It would be interesting as future work to find the measures and the preference vector $\mathbf{w} = [\mathrm{w}_1 \mathrm{w}_2 \cdots \mathrm{w}_M]$ that better solve the problem treated here. Other techniques for searching particular solutions from the set of non-dominated solutions can be found in [63].

Another suggestion for future work is to apply the methodology presented here to other dereverberation algorithms. In this work, the three considered algorithms belong to the speech enhancement family of dereverberation algorithms (see Section

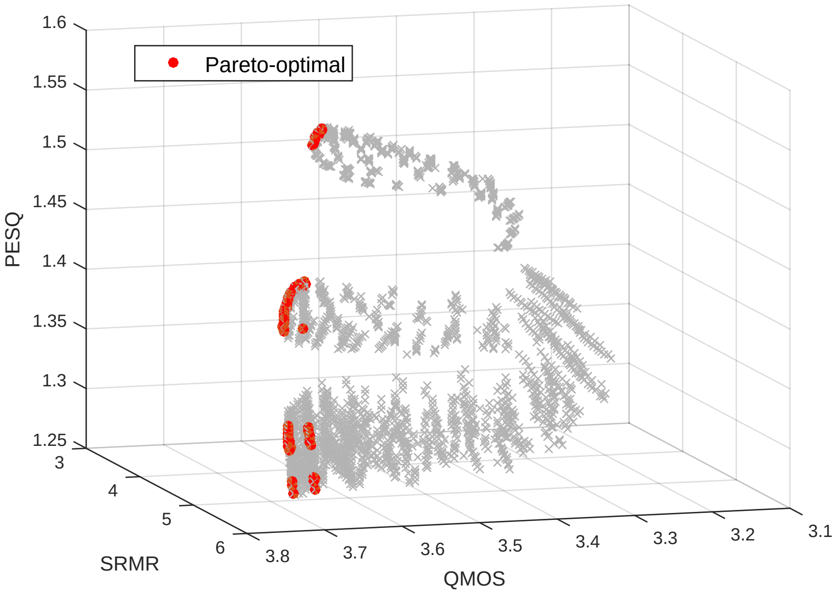


Figure 8.3: Example of feasible objective space using the PESQ, SRMR and Q_{MOS} measures.

2.3) and employ a single microphone. It would be interesting to apply the enhancement technique to different algorithms, such as those based on blind deconvolutions or beamforming, and with higher number of sensors.

Finally, it is also worth studying the problem of desreverberation and the solutions obtained here in the context of high-quality speech signals.

Bibliography

[1] DE LIMA, A. A., PREGO, T. D. M., NETTO, S. L., et al. "On the Quality-assessment of Reverberated Speech", *Speech Communication*, v. 54, n. 3, pp. 393–401, 2012.

[2] KINOSHITA, K., NAKTANI, T. *Speech Dereverberation using Linear Prediction.* In: NTT Technical Review 9 No.7, 2011.

[3] NAYLOR, P. A., GAUBITCH, N. D. *Speech Dereverberation.* London, Springer-Verlag, 2010.

[4] LOIZOU, P. *Speech Enhancement: Theory and Practice.* Taylor & Francis, 2007.

[5] ROBINSON, T., FRANSEN, J., PYE, D., et al. "Wsjcam0: A British English Speech Corpus For Large Vocabulary Continuous Speech Recognition". In: *Proceedings ICASSP 95*, pp. 81–84. IEEE, 1995.

[6] KINOSHITA, K., DELCROIX, M., YOSHIOKA, T., et al. "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech". In: *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4, 2013.

[7] "REVERB Challenge ASR evaluation tool (Baseline recognition system)". `http://reverb2014.dereverberation.com/download.html`.

[8] HABETS, E. A. P. *Single- and multi-microphone speech dereverberation using spectral enhancement.* Ph.D. dissertation, Technische Universiteit Eindhoven, 2007.

[9] SABINE, W. *Collected Papers on Acoustics.* Harvard University Press, 1922.

[10] RATNAM, R., JONES, D. L., WHEELER, B. C., et al. "Blind estimation of reverberation time", *The Journal of the Acoustical Society of America*, v. 114, n. 5, 2003.

[11] SCHROEDER, M. R. "New Method of Measuring Reverberation Time", *The Journal of the Acoustical Society of America*, v. 37, n. 3, 1965.

[12] ISO 3382:1997. "Acoustics - measurement of the reverberation time of rooms with reference to other acoustical parameters". 1997.

[13] RATNAM, R., JONES, D. L., O'BRIEN, W. D. "Fast algorithms for blind estimation of reverberation time", *IEEE Signal Processing Letters*, v. 11, n. 6, pp. 537–540, 2004.

[14] PREGO, T. D. M., DE LIMA, A. A., NETTO, S. L., et al. "A blind algorithm for reverberation-time estimation using subband decomposition of speech signals", *The Journal of the Acoustical Society of America*, v. 131, n. 4, 2012.

[15] HEINRICH LÖLLMANN, EMRE YILMAZ, M. J., VARY, P. "An Improved Algorithm for Blind Reverberation Time Estimation". In: *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC).*, pp. 1–4, 2010.

[16] FALK, T. H., ZHENG, C., CHAN, W. Y. "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech", *IEEE Transactions on Audio, Speech, and Language Processing*, v. 18, n. 7, pp. 1766–1774, 2010.

[17] DE M. PREGO, T., DE LIMA, A. A., ZAMBRANO-LÓPEZ, R., et al. "Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition". In: *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on*, pp. 1–5, 2015.

[18] JETZT, J. J. "Critical distance measurement of rooms from the sound energy spectral response", *The Journal of the Acoustical Society of America*, v. 65, n. 5, pp. 1204–1211, 1979.

[19] HABETS, E. A. P., GANNOT, S., COHEN, I. "Late Reverberant Spectral Variance Estimation Based on a Statistical Model", *IEEE Signal Processing Letters*, v. 16, n. 9, pp. 770–773, 2009.

[20] ALLEN, J. B., BERKLEY, D. A., BLAUERT, J. "Multimicrophone signal-processing technique to remove room reverberation from speech signals", *The Journal of the Acoustical Society of America*, v. 62, n. 4, 1977.

[21] WISDOM, S., POWERS, T., ATLAS, L., et al. "Enhancement of reverberant and noisy speech by extending its coherence". In: *Proceedings REVERB Challenge Workshop, Florence, Italy*, pp. 1–5, 2014.

[22] BRANDSTEIN, M. S., GRIEBEL, S. M. "Nonlinear, Model-Based Microphone Array Speech Enhancement". In: Gay, S. L., Benesty, J. (Eds.), *Acoustic Signal Processing for Telecommunication*, pp. 261–279, Boston, MA, Springer US, 2000.

[23] YEGNANARAYANA, B., MURTHY, P. S. "Enhancement of reverberant speech using LP residual signal", *IEEE Transactions on Speech and Audio Processing*, v. 8, n. 3, pp. 267–281, 2000.

[24] GRIEBEL, S. *A microphone array system for speech source localization, denoising and dereverberation*. Ph.D. dissertation, Harvard University, Cambridge, Massachusetts, 2002.

[25] GRIEBEL, S. M., BRANDSTEIN, M. S. "Microphone array speech dereverberation using coarse channel modeling". In: *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, v. 1, pp. 201–204 vol.1, 2001.

[26] YEGNANARAYANA, B., PRASANNA, S. R. M., RAO, K. S. "Speech enhancement using excitation source information". In: *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, v. 1, pp. I–541–I–544, 2002.

[27] GILLESPIE, B. W., MALVAR, H. S., FLORENCIO, D. A. F. "Speech dereverberation via maximum-kurtosis subband adaptive filtering". In: *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, v. 6, pp. 3701–3704 vol.6, 2001.

[28] WU, M., WANG, D. "A Two-Stage Algorithm for Enhancement of Reverberant Speech". In: *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, v. 1, pp. 1085–1088, 2005.

[29] NAKATANI, T., MIYOSHI, M., KINOSHITA, K. "Single-Microphone Blind Dereverberation". In: *Speech Enhancement*, pp. 247–270, Berlin, Heidelberg, Springer Berlin Heidelberg, 2005.

[30] LEBART, K., BOUCHER, J. M., DENBIGH, P. N. "A new method based on spectral subtraction for speech dereverberation", *Acta Acoust*, pp. 359–366, 2001.

[31] DE M. PREGO, T., DE LIMA, A. A., NETTO, S. L. "On the enhancement of dereverberation algorithms based on a perceptual evaluation criterion."

In: Bimbot, F., Cerisara, C., Fougeron, C., et al. (Eds.), *INTERSPEECH*, pp. 1360–1364. ISCA, 2013.

[32] GANNOT, S., MOONEN, M. "Subspace Methods for Multimicrophone Speech Dereverberation", *EURASIP Journal on Advances in Signal Processing*, v. 2003, n. 11, pp. 1–17, 2003.

[33] HUANG, Y., BENESTY, J., CHEN, J. "A Blind Channel Identification-Based Two-Stage Approach to Separation and Dereverberation of Speech Signals in a Reverberant Environment", *IEEE Transactions on Speech and Audio Processing*, v. 13, n. 5, pp. 882–895, 2005.

[34] SUBRAMANIAM, S., PETROPULU, A. P., WENDT, C. "Cepstrum-based deconvolution for speech dereverberation", *IEEE Transactions on Speech and Audio Processing*, v. 4, n. 5, pp. 392–396, 1996.

[35] TRIKI, M., SLOCK, D. T. M. "Delay and Predict Equalization for Blind Speech Dereverberation". In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, v. 5, p. V, 2006.

[36] DELCROIX, M., HIKICHI, T., MIYOSHI, M. "Precise Dereverberation Using Multichannel Linear Prediction", *IEEE Transactions on Audio, Speech, and Language Processing*, v. 15, n. 2, pp. 430–440, 2007.

[37] EVERS, C., HOPGOOD, J. R., BELL, J. "Acoustic models for online blind source dereverberation using sequential Monte Carlo methods". In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4597–4600, 2008.

[38] MIYOSHI, M., KANEDA, Y. "Inverse filtering of room acoustics", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 36, n. 2, pp. 145–152, 1988.

[39] FLANAGAN, J. L., JOHNSTON, J. D., ZAHN, R., et al. "Computer-steered microphone arrays for sound transduction in large rooms", *The Journal of the Acoustical Society of America*, v. 78, n. 5, 1985.

[40] GANNOT, S., BURSHTEIN, D., WEINSTEIN, E. "Signal enhancement using beamforming and nonstationarity with applications to speech", *IEEE Transactions on Signal Processing*, v. 49, n. 8, pp. 1614–1626, 2001.

[41] NISHIURA, T., NAKANURA, S., SHIKANO, K. "Speech enhancement by multiple beamforming with reflection signal equalization". In: *Acoustics,*

Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on, v. 1, pp. 189–192 vol.1, 2001.

[42] REMEZ, R., RUBIN, P., PISONI, D., et al. "Speech perception without traditional speech cues", *Science*, v. 212, n. 4497, pp. 947–949, 1981.

[43] "Methods for subjective determination of transmission quality. Recommendation P.800, International Telecommunications Union (ITU-T)". 1996.

[44] ALLEN, J. B. "Effects of small room reverberation on subjective preference", *The Journal of the Acoustical Society of America*, v. 71, n. S1, 1982.

[45] NAYLOR, A. P., HABETS, P. E. A., WEN, Y.-C. J., et al. "Models, Measurement and Evaluation". In: Naylor, A. P., Gaubitch, D. N. (Eds.), *Speech Dereverberation*, pp. 21–56, London, Springer London, 2010.

[46] J.Y.C.WEN, NAYLOR, P. "An evaluation measure for reverberant speech using tail decay modelling". In: *Proceedings of the European Signal Processing Conference (EUSIPCO'06), Florence, Italy,*, pp. 1–4, 2006.

[47] "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs". Recommendation P.862, International Telecommunications Union (ITU-T)". 2000.

[48] "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs". Recommendation P.862.2, International Telecommunications Union (ITU-T)". 2007.

[49] BISPO, B. C., ESQUEF, P. A. A., BISCAINHO, L. W. P., et al. "EW-PESQ: A Quality Assessment Method for Speech Signals Sampled at 48 kHz", *Journal of the Audio Engineering Society*, v. 58, n. 4, pp. 251–268, 2010.

[50] WANG, S., SEKEY, A., GERSHO, A. "An objective measure for predicting subjective quality of speech coders", *IEEE Journal on Selected Areas in Communications*, v. 10, n. 5, pp. 819–829, Jun 1992.

[51] HANSEN, J. H. L., PELLOM, B. L. "An Effective Quality Evaluation Protocol For Speech Enhancement Algorithms". In: *Proceedings of the International Conference on Speech and Language Processingop*, pp. 2819–2822, 1998.

[52] TRIBOLET, J., NOLL, P., MCDERMOTT, B., et al. "A study of complexity and quality of speech waveform coders". In: *Acoustics, Speech, and Signal*

Processing, IEEE International Conference on ICASSP '78., v. 3, pp. 586–590, 1978.

[53] S. QUACKENBUSH, T. B., CLEMENTS, M. *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[54] RIX, A. W. "Perceptual speech quality assessment - a review". In: *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, v. 3, pp. iii–1056–9 vol.3, 2004.

[55] GRAY, P., HOLLIER, M. P., MASSARA, R. E. "Non-intrusive speech-quality assessment using vocal-tract models", *IEE Proceedings - Vision, Image and Signal Processing*, v. 147, n. 6, pp. 493–501, 2000.

[56] CHEN, G., PARSA, V. "Nonintrusive speech quality evaluation using an adaptive neurofuzzy inference system", *IEEE Signal Processing Letters*, v. 12, n. 5, pp. 403–406, 2005.

[57] JIN, C., KUBICHEK, R. "Vector quantization techniques for output-based objective speech quality". In: *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, v. 1, pp. 491–494 vol. 1, 1996.

[58] PICOVICI, D., MAHDI, A. E. "Output-based objective speech quality measure using self-organizing map". In: *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, v. 1, pp. I–476–I–479 vol.1, 2003.

[59] KIM, D.-S., TARRAF, A. "Perceptual model for non-intrusive speech quality assessment". In: *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, v. 3, pp. iii–1060–3 vol.3, 2004.

[60] STEENEKEN, H. J. M., HOUTGAST, T. "A physical method for measuring speech-transmission quality", *The Journal of the Acoustical Society of America*, v. 67, n. 1, 1980.

[61] HAZRATI, O., LOIZOU, P. C. "Tackling the Combined Effects of Reverberation and Masking Noise Using Ideal Channel Selection", *Journal of Speech, Language, and Hearing Research*, v. 55, n. 2, 2012.

[62] HOLUBE, I., KOLLMEIER, B. "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception

model", *The Journal of the Acoustical Society of America*, v. 100, n. 3, 1996.

[63] DEB, K. *Multi-Objective Optimization Using Evolutionary Algorithms.* New York, NY, USA, John Wiley & Sons, Inc., 2001.

[64] KINOSHITA, K., DELCROIX, M., GANNOT, S., et al. "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research", *EURASIP Journal on Advances in Signal Processing*, v. 2016, n. 1, pp. 1–19, 2016.

[65] LINCOLN, M., MCCOWAN, I., VEPA, J., et al. "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): specification and initial experiments". In: *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pp. 357–362, 2005. doi: 10.1109/ASRU.2005.1566470.

[66] YOUNG, S., EVERMANN, G., GALES, M., et al. *The HTK Book.* Cambridge University Engineering Department, 2006.

[67] DAVIS, S., MERMELSTEIN, P. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 28, n. 4, pp. 357–366, 1980.

[68] PABLO CANCELA, E. L., ROCAMORA, M. "Fan Chirp Transform for Music Representation". In: *Proceedings International Conference On Digital Audio Effects (DAFx)*, pp. 1–8, 2010.

[69] EPHRAIM, Y., MALAH, D. "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 32, n. 6, pp. 1109–1121, 1984.

[70] COHEN, I. "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator", *IEEE Signal Processing Letters*, v. 9, n. 4, pp. 113–116, 2002.

[71] COHEN, I. "Speech Enhancement". cap. From Volatility Modeling of Financial Time-Series to Stochastic Modeling and Enhancement of Speech Signals, pp. 97–113, Berlin, Heidelberg, Springer Berlin Heidelberg, 2005.

[72] RIBAS, D., CRESPO, S., CALVO, J. R. “Single channel speech enhancement based on zero phase transformation in reverberated environments”. In: *Proceedings REVERB Challenge Workshop, Florence, Italy*, pp. 1–5, 2014.