12-16-2019 10:00 AM

Objective Estimation of Tracheoesophageal Speech Quality

Yousef S Ettomi Ali
The University of Western Ontario

Supervisor
Parsa, Vijay
The University of Western Ontario

Graduate Program in Electrical and Computer Engineering
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Abstract

Speech quality estimation for pathological voices is becoming an increasingly important research topic. The assessment of the quality and the degree of severity of a disordered speech is important to the clinical treatment and rehabilitation of patients. In particular, patients who have undergone total laryngectomy (larynx removal) produce Tracheoesophageal (TE) speech. In this thesis, we study the problem of TE speech quality estimation using advanced signal processing approaches. Since it is not possible to have a reference (clean) signal corresponding to a given TE speech (disordered) signal, we investigate in particular the non-intrusive techniques (also called single-ended or blind approaches) that do not require a reference signal to deduce the speech quality level.

First, we develop a novel TE speech quality estimation based on some existing double-ended (intrusive) speech quality evaluation techniques such as the Perceptual Evaluation Speech Quality (PESQ) and Hearing Aid Speech Quality Index HASQI. The matching pursuit algorithm (MPA) was used to generate a quasi-clean speech signal from a given disordered TE speech signal. Then, by adequately choosing the parameters of the MPA (atoms, number of iterations,...etc) and using the resulting signal as our reference signal in the intrusive algorithm, we show that the resulting intrusive algorithm correlates well with the subjective scores of two TE speech databases.

Second, we investigate the extraction of low complexity auditory features for the evaluation of speech quality. An 18-th order Linear Prediction (LP) analysis is performed on each voiced frame of the speech signal. Two evaluation features are extracted corresponding to higher order statistics of the LP coefficients and the vocal

tract model parameters (cross-sectional tubes areas). Using a set of 35 TE speech samples, we perform forward stepwise regression as well as K-fold cross validation to select the best sets of features that are used in each of the regression models. Finally, the selected features are fitted to different support vector regression models yielding high correlations with subjective scores.

Finally, we investigate a new approach for the estimation of the quality of TE speech using deep neural networks (DNNs). A synthetic dataset that consists of 2173 samples was used to train a DNN model that was shown to predict the TE voice quality. The synthetic dataset was formed by mixing 53 normal speech samples with modulated noise signals that had a similar envelope to the speech samples, at different speech-to-modulation noise ratios. A validated instrumental speech quality predictor was used to quantify the perceived quality of speech samples in this database, and these objective quality scores were used for training the DNN model. The DNN model was comprised of an input layer that accepted sixty relevant features extracted through filterbank and linear prediction analyses of the input speech signal, two hidden layers with 15 neurons each, and an output layer that produced the predicted speech quality score. The DNN trained on the synthetic dataset was subsequently applied to four different databases that contained speech samples collected from TE speakers. The DNN-estimated quality scores exhibited strong correlation with the subjective ratings of the TE samples in all four databases, thus it shows a strong robustness compared to those speech quality metrics developed in this thesis or those from the literature.

Summary for Lay Audience

Speech quality estimation is a multi-dimensional perceptual phenomenon that encompasses attributes such as clarity, pleasantness, and naturalness of speech. There is a necessity to estimate the quality of pathological voices due to its clinical importance, especially for people who have undergone total laryngectomy (larynx removal). Speech coming out of people who undergo such surgery is called Tracheoesophageal (TE) speech. This thesis aims to assess the quality of TE speech using speech quality metrics that incorporate digital signal processing and machine learning algorithms.

In the first contribution of this study, a novel TE speech quality estimation metric is developed using intrusive techniques. Intrusive techniques are those methods that need a clean reference audio signal to measure the quality of the signal that is being evaluated. The obtained automated model is found to have high similarity in terms of performance to the subjective human evaluation of speech audio signals.

In the second part of this study, the use of non-intrusive metrics that do not need a clean reference signal to evaluate the quality of speech signals is investigated. Linear prediction features from the speech signal are fed to a stepwise regression model to predict the quality of the speech records. Moreover, another machine learning algorithm, support vector regression (SVR) is used to extract the quality evaluation metrics from these prediction coefficients. The obtained quality metrics are found to be highly correlated with individual subjective scores.

Finally, the third part of this study investigates the use of artificial deep neural networks (DNN), a state-of-the-art machine learning technique, in predicting the quality of TE speech records. The DNN-estimated quality scores exhibited a strong correlation with the subjective ratings of the TE samples in all four databases, thus it

shows strong robustness compared to those speech quality metrics developed in this thesis or those from the literature.

Dedicated to the memory of my parents Aysha Almbrook and Souleimen Ettomi

Acknowledgment

I would like to express my deepest gratitude to my supervisor, Dr. Vijay Parsa, for his patienc guidance, support, encouragement and advice throughout my the time I have spent at University of Western Ontario. I have been extremely lucky to have a supervisor who cared so much about my progress at work, and who responded to my questions and queries so promptly which helped me to smoothly conduct my research and write my thesis.

I am grateful to Dr. Philip C. Doyle, who have helped me to complete several parts of this research and has provided with the experimental speech databases for the validation of my results. I also thankful to my friend, colleague and collaborator, Dr. Soulaimane Berkane, with whom I shared countless discussions in this research area. Special thanks to Professor Scott Adams who have shared with me insightful thoughts on speech acoustics and production.

I would also like to sincerely thank Dr. Sreeraman Rajan, Dr. Samuel Asokanthan, Dr. Raveendra K. Rao and Dr. Abdelkader Ouda for taking the time to serve as my PhD thesis examiners and for their constructive comments and feedback.

Last but not the least, I would like to extend my sincere appreciation to my family, to my wife and kids, for their love, support and prayers in all my pursuits.

Contents

Abstract **i**

Lay summary **iii**

Acknowledgment **vi**

List of Figures **xi**

List of Tables **xv**

List of Abbreviations **xvii**

1 Introduction **1**

1.1 General Introduction 1

1.2 Human Speech Production System 3

1.2.1 Pulmonary Pressure 4

1.2.2 Phonation 4

1.2.3 Vowels and Consonants 5

1.3 Acoustic Theory of Speech Production 6

1.4 Tracheoesophageal Disordered Speech 7

1.5 Disordered Speech Quality Estimation 8
1.6 Thesis Scope 15
1.7 Thesis Contributions 16
1.8 Thesis Outline 19

2 Background and Preliminaries **21**
2.1 Signal Processing Tools 21
2.1.1 Forward Stepwise Regression 21
2.1.2 K-folds cross-validation 22
2.1.3 Mean Squared Error 22
2.1.4 Linear Prediction 23
2.1.5 Pearson correlation 25
2.1.6 Spearman rank correlation 26
2.1.7 Sigmoidal mapping function 26
2.1.8 Higher-Order-Statistics 26
2.1.9 The ITU-T P.862 standard 27
2.1.10 The Hearing Aid Speech Quality Index (HASQI) 30
2.2 The ITU-T P.563 Standard 33
2.3 TE speech databases 36
2.4 Conclusion 37

3 Disordered speech quality estimation using the matching pursuit algorithm **39**
3.1 Introduction 39
3.2 Quasi-Reference Perceptual Speech Quality Estimation 40

3.2.1 Matching Pursuit Algorithm 41
3.2.1.1 Energy Capture Ratio (ECR) feature 43
3.2.2 Proposed speech quality estimation features 44
3.2.2.1 MPA-PESQ feature 46
3.2.2.2 MPA-HASQI feature 48
3.2.2.3 Residual Rate Feature (RRF) 49
3.3 Evaluation Method 52
3.4 Results 53
3.5 Discussions 58
3.6 Conclusion 61

4 Low Complexity Disordered Speech Quality Estimation 62
4.1 Introduction 62
4.2 Speech quality evaluation method 63
4.2.1 Pitch Period Estimation and Voiced Frames Extraction 64
4.2.2 Linear Prediction Analysis 65
4.2.2.1 Cepstral Coefficients 65
4.2.2.2 LP Residual 66
4.2.3 Vocal Tract Modelling 66
4.2.4 Features Extracted 68
4.2.4.1 Higher Order Statistics 68
4.2.4.2 Vocal Tract Parameters 71
4.3 Speech database 72
4.4 Results and discussions 72
4.5 Conclusion 78

5 Deep Learning-Based Quality Assessment for Tracheoesophageal Speech 81

5.1 Introduction 81

5.2 Speech Quality Evaluation Method 82

5.2.1 Linear Prediction-Based Higher Order Statistics 83

5.2.2 Vocal Tract Parameters 83

5.2.3 Mel Frequency Cepstrum Coefficients (MFCCs) 84

5.2.4 Gammatone Frequency Cepstrum Coefficients (GFCCs) 85

5.3 Deep Learning 86

5.4 Methodology 87

5.4.1 Artificial Subjective Dataset 88

5.4.2 Feature Selection and Reduction 91

5.4.3 Deep Neural Network (DNN) Training 92

5.5 Results 94

5.5.1 TE speech dataset 94

5.5.2 Evaluation and Performance 95

5.6 Discussion 97

5.7 Conclusion 104

6 Conclusion 105

6.1 Summary 105

6.2 Future Directions 106

A CAPE-V 108

List of Figures

1.1 Clinical setting for disordered speech quality measurement 2
1.2 Different stages in human speech production process. 3
1.3 Human speech production system anatomy 5
1.4 Engineering model of speech production. 6
1.5 Tracheoesophageal Speech Production 9
1.6 Perceived versus CSID-estimated severity ratings 15

2.1 An illustration of the linear prediction process. 25
2.2 Block diagram of the Perceptual Speech Quality Evaluation (PESQ) computational procedure. 28
2.3 Diffrent types of indices used in the HASQI process. 31
2.4 ITU-T P.563 schematic diagram. 34
2.5 Pitch synchronous vocal tract model and LPC analysis 35
2.6 Full-reference quality assessment model in P.563 36

3.1 Schematic of the proposed quasi-reference intrusive speech quality estimation algorithm 40
3.2 Illustration of the Gaussian window with different values of a and b. . 42

3.3 Illustration of the matching pursuit algorithm. a) Input (original) signal b) The used time-frequency atom c) Residual signal after removal of energy represented by the time-frequency atom.) 44
3.4 Energy capture ratio for normal and disordered speech samples. . . . 45
3.5 An original speech sample from the first Database D1. 46
3.6 Reconstruction of the speech data sample in Figure 3.5 using the MPA at different iterations. 46
3.7 Schematic of the MPA-PESQ quality feature. 48
3.8 Schematic of the MPA-HASQI quality feature. 49
3.9 Derived matching pursuit features ECR and RRF based on the energy capture of the original and reconstructed signal. 50
3.10 Logarithm of the residual energy ratio (solid) and its corresponding linear approximation (dashed) obtained by a least-mean square method. 51
3.11 Correlation results for the voice samples of the all TE speaker databases D1, D2, D3, and D4 54
3.12 Subjective severity score for database D_1 and MPA-PESQ feature and MPA-HASQI feature. 56
3.13 Subjective quality score for database $D2$ and MPA-PESQ, MPA-HASQI 57
3.14 Subjective severity score for database $D3$ and MPA-PESQ, MPA-HASQI 57
3.15 Subjective quality score for database $D4$ and MPA-PESQ, MPA-HASQI 58

3.16 MPA-PESQ feature for the voice samples of the all TE speaker database with respect to the number of iterations of the MPA. The maximum score and minimum score samples are plotted in bold. 59
3.17 MPA-HASQI feature for the voice samples of the all TE speaker database with respect to the number of iterations of the MPA. The maximum score and minimum score samples are plotted in bold. 60

4.1 The proposed speech quality algorithm. 63
4.2 Pitch period estimation and voiced frames extraction method using the autocorrelation method. 65
4.3 Illustration of the vocal tract uniform-cross-sectional-area tube model 69
4.4 Average value of LP coefficients for each voiced TE speech frame. . . 70
4.5 Feature selection from the HOS statistics group 73
4.6 Feature selection from VTP Parameters group 75
4.7 Scatter plot of subjective scores against the objective scores derived from the VTP parameters-based model. 76
4.8 Scatter plot of subjective scores against the objective scores derived from the HOS statistics-based model. 77

5.1 Proposed algorithm for disordered speech quality estimation. The DNN network is pre-trained using a custom synthetic dataset and then is used to map the extracted speech features (MFCC, GFCC, HOS, VTP) into a quality score. 88

5.2 We have generated an artificial disordered speech dataset (2173 samples) using clean normal speech samples and speech like noise. The HASQI metric was then used to predict the perceived quality of each deliberately distorted speech signal. We compared between two types of speech like noise: speech shaped noise (SNN) and the Modulated Noise Reference Unit (MNRU) noise. 89
5.3 The HASQI values for the two artificially generated speech datasets . 90
5.4 Out-of-sample mean square error during the feature reduction process. A total of 60 features is a reasonable choice that optimizes the out-of-sample MSE while guaranteeing the smallest number of features possible. 93
5.5 DNN training results for both artificially generated datasets. 94
5.6 Subjective severity score for database D_1 and predictive quality score based on SMNRU and SSN. 99
5.7 Subjective severity score for database D_2 and predictive quality score based on SMNRU and SSN. 100
5.8 Subjective quality score for database D_3 and predictive quality score based on SMNRU and SSN. 101
5.9 Subjective severity score for database D_4 and predictive quality score based on SMNRU and SSN. 102
5.10 Normalized MFCC and GFCC coefficients. 103

List of Tables

1.1 Example of the Consensus Auditory Perceptual Evaluation-Voice (CAPE-V) scale score sheet. 11

2.1 Statistics of the subjective scores for the three reference TE speech datasets. 37

3.1 Correlation values for different objective metrics for database D_1. . . 53
3.2 Correlation values for different objective metrics for database D_2. . . 55
3.3 Correlation values for different objective metrics for database $D3$. . . 55
3.4 Correlation values for different objective metrics for database D_4. . . 55

4.1 High-order statistics (HOS) features. 70
4.2 Forward stepwise regression results. 74
4.3 Selected features for each model. 75
4.4 Correlation values of the proposed objective metrics. 76
4.5 Comparison of the correlation values obtained using different quality estimation methods. 79

5.1 Statistics of the HASQI scores for the two synthetic datasets. 91
5.2 Correlation values for different objective metrics for database D_1. . . 97

5.3 Correlation values for different objective metrics for database D_2. . . 97

5.4 Correlation values for different objective metrics for database D_3. . . 98

5.5 Correlation values for different objective metrics for database D_4. . . 98

List of Abbreviations

ADSD	Adductor Spasmodic Dysphonia
ADSV	Analysis of Dysphonia in Speech and Voice
ANN	Artificial Neural Network
BM	Basilar membrane
BVFLs	Benign Vocal Fold Lesions
CAPE-V	Consensus Auditory-Perceptual Evaluation of Voice
CASA	Computational Auditory Sense Analysis
CPPs	Smoothed Cepstrum Peak Prominence
CSID	Cepstral Spectral Index of Dysphonia
DNNs	deep neural networks
ECR	Energy Capture Ratio
ERB	Equivalent Rectangular Bandwidth
FFT	Fast Fourier Transform

FSR	Forward Stepwise Regression
GFCCs	Gammatone Frequency Cepstrum Coefficients
GRBAS	Grade, Roughness, Breathiness, Asthenicity, and Strain
HASQI	Hearing Aid Speech Quality Index
HI	Hearing-Impaired
HNR	Harmonics-to-Noise-Ratio
HOS	higher-order statistics
IRS	Intermediate Reference System
ITU	International Telecommunication Union
LCDSQE	Low Complexity Disordered Speech Quality Estimation
LCQA	Low-Complexity Quality Assessment
LP	Linear Prediction
LTAS	Long-term average spectra
MFCCs	Mel Frequency Cepstrum Coefficients
MNRU	Modulated Noise Reference Unit
MOS	Mean Opinion Score
MP	Matching Pursuit

MPA	Matching Pursuit Algorithm
MSE	Mean Squared Error
NH	Normal Hearing
PESQ	Perceptual Evaluation Speech Quality
PESQ	Perceptual Evaluation of Speech Quality
PMTD	Primary Muscle Tension Dysphonia
RAP	Relative Average Perturbation
RRF	Residual Rate Feature
SIG3	Special Interest Group
SMNRU	Speech Modulated Noise Reference Unit
SNN	speech shaped noise
SR	Stepwise Regression
TE	Tracheoesophageal
UVFP	Unilateral Vocal fold Paralysis
VB	Voice Breaks

Chapter 1

Introduction

1.1 General Introduction

Speech communication plays a vital role in our lives and disorders affecting speech communication can have significant impact on psychological, physical, and financial well-being. Speech and voice disorders are categorized according to their causes, syndromes and treatment. These can occur due to physiological or psychological disorders, accidents, misuse of voice, or surgery affecting the vocal folds. In addition, a variety of diseases and medical complications can cause speech and voice abnormalities. Early identification of the speech and/or voice disorder, followed by proper intervention and monitoring are crucial for recuperation and improved quality of life [1]. Clinicians and researchers working with persons with speech/voice disorders employ evidence-based assessment and treatment strategies, and strive to demonstrate positive and objective outcomes associated with their interventions [2]. Figure 1.1 shows clinical setting for disordered speech quality measurement. Typically, speech/voice

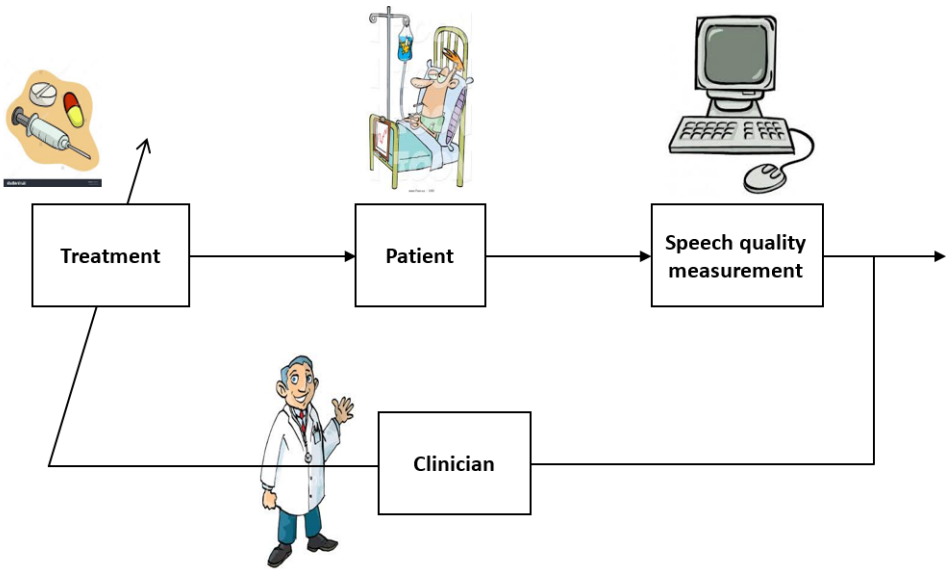

Figure 1.1: Clinical setting for disordered speech quality measurement

disorders are assessed through different modalities including perceptual, acoustic, aerodynamic, and endoscopic imaging techniques [3]. With the advent of relatively inexpensive personal computers, low cost analysis software, and increased availability of digital audio recording systems, acoustic assessment has become an increasingly popular option for tracking intervention outcomes. Advancement is being made on the improvement of acoustic analysis algorithms that are more robust for analysing disordered voices [4, 5, 6, 7]. This thesis concentrates on the perceptual and acoustic methods of disordered speech quality assessment. Novel acoustic and perceptual features are proposed for a class of disordered speech.

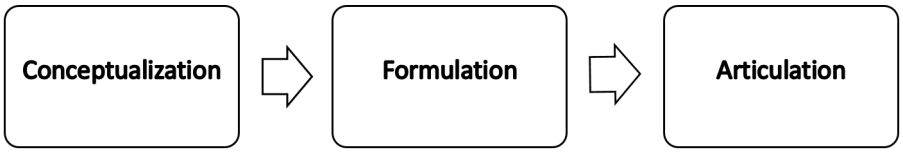

Figure 1.2: Different stages in human speech production process.

1.2 Human Speech Production System

Linguistically, speech is the ability to express thoughts and feelings through articulation of sounds. Speech production is the process that translates brain thoughts into hearble speech. It involves many organs, muscles and intermediate complex processes. Roughly speaking, speech is produced through three major stages (levels): conceptualization, formulation, and articulation [8]. In the first stage, called also conceptual preparation, the human intention to generate a speech links the desired concept to the particular spoken words to be expressed [9]. The second stage (called formulation) allows to create a linguistic form for the desired message to be spoken. This includes grammatical, phonetic and morpho-phonological encoding [9]. The last stage in the speech production process is the articulation which is the mechanical execution of the linguistic form generated from the formulation stage, see Figure 1.2. Articulating a speech involves organs such as the lungs, glottis, larynx, tongue, lips, jaw and other parts of the vocal apparatus. These organs can be grouped into three main parts which acts to produce the pulmonary pressure, the phonation and the vowels and consonants, see Figure 1.3.

1.2.1 Pulmonary Pressure

The lungs provide the main source of excitation (energy source) in the speech production process. When exhaling air, the volume of the chest cavity is reduced which causes the lung air pressure to increase. This increase in pressure causes air to flow through the trachea (also called windpipe) into the larynx [10].

1.2.2 Phonation

The air flow generated by the pulmonary pressure is made audible when set into vibration by the activity of the larynx. The *larynx*, composed of muscles, ligaments and cartilages, controls the function of the *vocal folds* (also called vocal chords). The vocal folds are two membranes (two masses of ligament) stretching horizontally from the posterior to the interior of the larynx. The opening between the two vocal folds is called the *glottis*. The vocal folds are typically 15 mm long in men and about 13 mm in women. By means of various muscle contractions, the vocal folds can be varied in length and thickness and positioned in various configurations [11] . These contractions cause a change in the characteristics of the quasi-periodic airflow which result in a change of the waveshape of the glottal air pulses, including their duration and amplitude, which lead to changes in perceived pitch and loudness. The perceived pitch is the physical aspect corresponding to the fundamental frequency, denoted F0, of the speech signal generated by the vocal folds vibration.

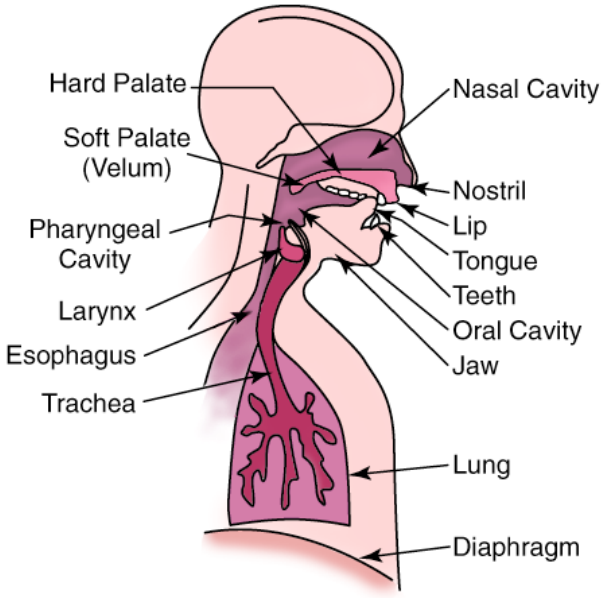

Figure 1.3: Human speech production system anatomy [12].

1.2.3 Vowels and Consonants

The sound produced by the larynx is further shaped (or modified) when passing through the *vocal tract*. The vocal tract consists of two main parts: the nasal cavity and the oral cavity. The oral cavity extends from the larynx to the lips while the nasal cavity is coupled to the oral cavity through the *velum*. The oral cavity consists of the tongue, teeth, lips and jaws which are known as the *articulators*. Altering these articulators resuls in different shapes of the oral tract which allows different filtering of the phonation sound [10]. Further details about the mathematical modelling of the acoustic speech production are detailed in the next section.

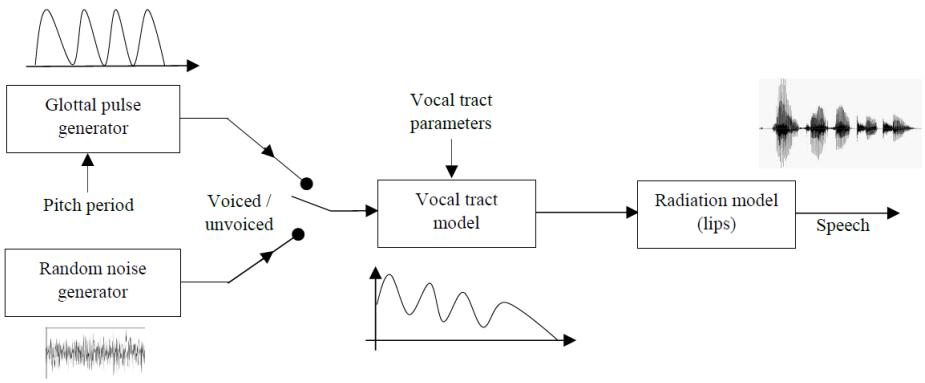

Figure 1.4: Engineering model of speech production.

1.3 Acoustic Theory of Speech Production

The acoustic characteristics of speech are usually modelled as a sequence of source, vocal tract filter, and radiation characteristics [10]. Depending on their state, the vocal folds provide excitation, which can be periodic or aperiodic, to the vocal tract. *Voiced* sounds, such as vowels, are produced when the vocal folds vibrate and provide periodic source excitation. When the source of excitation is aperiodic (e.g. random noise) the produced sound is *unvoiced*. The vocal tract acts as a filter that spectrally shapes the input excitation provided by the vocal folds to produce various sounds. This sound is further filtered by the effect of sound radiation at the level of the lips. These three stages define the engineering model of speech production, see Figure 1.4.

The vocal tract can be modelled as an acoustic tube with resonances, called formants, and antiresonances. Moving the articulators of the vocal tract alters the shape of the acoustic tube which results in changes in the frequency response (i.e. changes in the formants as well). The wavesound generated by the larynx is considered as an input to the acoustic tube where certain frequencies are attenuated while others

(near formants) are amplified. In general, the acoustic tube corresponding to the vocal tract is often taken as a chain of cylinders with different cross sectional areas (tubes connected in series). Note that this is a simplified model and the actual vocal tract involves much more complex shapes and configurations.

1.4 Tracheoesophageal Disordered Speech

Some patients are diagnosed with laryngeal cancer which often requires the surgical removal of the whole larynx or part of it. Removal of the entire larynx is termed total laryngectomy. In these circumstances, the speech production system is gravely damaged due to the removal of one of its important organs, which is the larynx. Therefore, alternative methods for voice generation are required. In a tracheoesophageal puncture procedure (surgery), a hole is created between the trachea and the esophagus (the tubal pathway between the throat and the stomach). The trachea is, subsequently, brought forward to the front of the neck and the individual will subsequently exhale and inhale from the neck. Further, because the larynx is decoupled from the upper trachea, the residual trachea is brought forward to the anterior midline neck where an open airway will exist for the remainder of the individual's life. A voice prosthesis is inserted into the puncture which keeps food out of the trachea but lets air into the esophagus for tracheoesophageal (TE) speech. Since early works by [13] this technique has became the international standard for voice restoration after total laryngectomy.

The speech produced by the TE voice production mechanism has often a lower quality compared to normal speech since the sound source is abnormal and contains

different anatomical asymmetries. TE speech is, generally, characterized by lowered frequency, normal or slightly greater than normal intensity, and because of access to the large volume of pulmonary air, generally normal temporal features (rate of speech) when compared to normal speakers [14]. However, the overall sound quality of TE speech is best described as highly aperiodic, rough, and noisy. Additionally, depending on the surgery and the length of treatment time with a speech pathologist, considerable variability across TE speakers does exist [15, 16].

1.5 Disordered Speech Quality Estimation

Disordered speech quality can be asssessed through perceptual (also termed *subjective*) and acoustic (also labeled as *objective*) measurements. Subjective perception of speech quality is a complex psychoacoustic phenomena that incorporates the interaction of many processes within human audition. The topic of speech quality perception is widely researched in telecommunications, audio recording & broadcasting, and architectural acoustics fields [18]. For example, in telecommunications, the Mean Opinion Score (MOS) is used to benchmark new transmission/reception or speech coding/decoding technologies [19]. In this subjective testing procedure, a panel of listeners rate the quality of speech sample under test on an integer scale of 1 (“bad" quality) to 5 (“excellent" quality), and the average of these ratings is reported as the perceived quality of the test speech sample. In the realm of subjective assessment of disordered speech/voice, the term *auditory-perceptual evaluation* is commonly used [20], and two of the more standard subjective assessment procedures include the GRBAS (Grade, Roughness, Breathiness, Asthenia, Strain) scale and the CAPE-V. In the

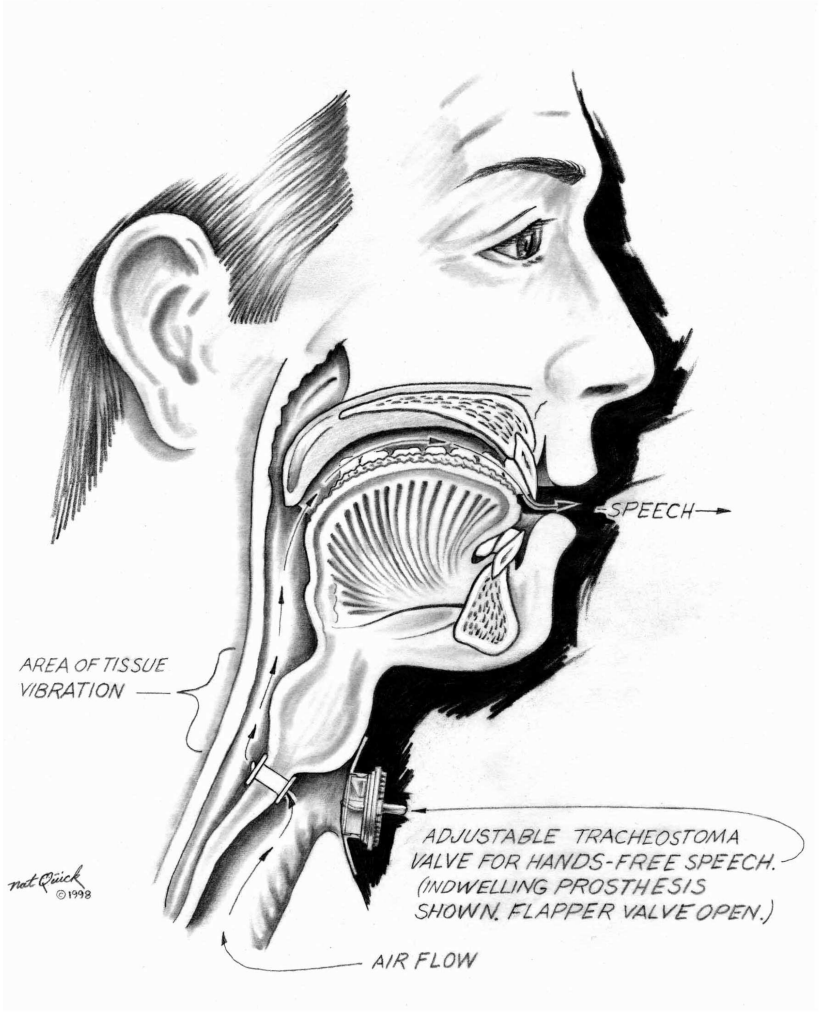

Figure 1.5: Tracheoesophageal Speech Production [17]

GRBAS procedure, the clinician rates the perceived quality along the Grade, Roughness, Breathiness, Asthenicity, and Strain dimensions on a scale of 0–3 [21]. More recently, the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) has been developed by the Special Interest Group (SIG3) of the American Speech-Language-Hearing Association as a standard clinical protocol for voice quality assessment[22]. The CAPE-V protocol specifies six quality features to be evaluated consistently and these are: overall severity, roughness, breathiness, strain, pitch and loudness (see Appendix A for a sample CAPE-V form). Karnell *et al.* (2007) published a preliminary report comparing the reliability of clinician-based auditory-perceptual judgments using the CAPE-V to those made with the GRBAS voice rating scheme. Karnell *et al.* found that both scales resulted in high interrater reliability while suggesting that the CAPE-V may offer "more sensitivity to small differences within and among patients than the GRBAS scale" [23].

While subjective measurements of disordered speech quality are preferred because of "greater intuitive meaning and shared reality among listeners" [20], they are influenced by factors such as listener experience, listeners' understanding of the rating scale, and the type of voice sample [20] all of which affect the reliability of subjective scores. In contrast to subjective voice quality estimation schemes, the objective methods use an *algorithm*, which replaces the listener panel, to compute the quality score from a given speech sample. Objective methods assesses speech quality through the use of physical characteristics of the speech signal and appropriate computational models [18]. Acoustic analysis methods have been used clinically to differentiate normal from abnormal voices, to evaluate the relative effectiveness of different treatment approaches, and to track progress in voice therapy.

Criteria	Mildly Deviant	Moderately Deviant	Severely Deviant	Score /100
Overall Quality		✓		63
Roughness			✓	30
Breathiness		✓		55
Strain	✓			88
Pitch			✓	25
Loudness		✓		60
Average Score				**53.5**

Table 1.1: Example of the Consensus Auditory Perceptual Evaluation-Voice (CAPE-V) scale score sheet.

There exist many approaches to the acoustic measurement of voice function. The most common are the time-based perturbation measures such as jitter and shimmer and their variants [24, 25]. Jitter and shimmer are acoustic characteristics of voice signals, and they are quantified as cycle-to-cycle variations of fundamental frequency and waveform amplitude, respectively. Fundamental frequency is determined physically by the vocal fold vibrations per second, and jitter represent the variations that occur in the fundamental frequency. The high amount of jitter is a consequence of erratic vibratory patterns due to loss of control of the vibratory system. It results in a voice with roughness that is usually perceived in the recordings of pathological voices. Therefore, a reliable estimation of jitter can be used to discriminate between healthy and dysphonic speakers [5]. There are three commonly used metrics to quantify jitter: absolute jitter, period (relative) jitter and Relative Average Perturbation

(RAP) jitter. Absolute jitter is the cycle-to-cycle variation of fundamental frequency, i.e the average absolute difference between consecutive periods:

$$\text{Relative jitter} = \frac{\frac{1}{n-1}\sum_{i=1}^{n-1}|T_i - T_{i-1}|}{\frac{1}{n}\sum_{i=1}^{n-1}T_i}, \tag{1.1}$$

where T_i is the length of the i^{th} cycle in ms and n is the number of cycles in the speech signal. Relative jitter is the average absolute difference between consecutive periods divided by the average period. RAP jitter is defined as, the average absolute difference between a period and the average of it and its two neighbours divided by the average period [24]. Shimmer, on the other hand, is an acoustics characteristic of voice signals that represents the cycle to cycle variation of waveform amplitude. Amplitude perturbations are similar to the frequency perturbation measures in the way that they attempt to quantify the short-term instability of the speech signal. The values of jitter and shimmer above a certain threshold are considered being related to pathological voices, which are usually perceived by humans as breathy, rough or hoarse voices.

The acoustic features such as jitter and shimmer are based on the assumption that the patient is attempting a relatively steady pitch and loudness production. Therefore, voice samples used for perturbation analyses are typically sustained vowels. However, in many clinical applications, it is important to analyse the speech quality in the case of continuous (connected) speech signal for different reasons. For instance, time-based acoustic measures such as jitter and shimmer rely on exact demarcation of the cycle-to-cycle boundaries in the acoustic waveform. This boundary detection is generally not reliable for non-periodic voices that characterize dysphonic individuals

[26, 27]. Also, sustained vowel-based measurements do not take into account some highly relevant vocal function attributes such as voice onset and termination, voice breaks, and variations in pitch and amplitude. In fact, several studies have found that measurements obtained from continuous speech may better predict voice quality than measurements from sustained vowels [28, 29, 30, 31]. For instance, Parsa *et al.* [5] found that using jitter as a classification index resulted in an accuracy of only 68% on a database of 53 normal 175 pathological talkers.

Relatively few studies have investigated the effectiveness of extracting acoustical features from continuous speech samples. The author [32] extended the Harmonic-to-Noise-Ratio (HNR) concept to continuous speech and reported a 5.6% error rate in classifying pathological talkers. Qi *et al.* [32, 33] proposed an SNR estimation technique based on Linear Prediction (LP) modelling and reported a correlation of 0.78 between subjective quality ratings and LP-based SNR. This algorithm has been extended in [34] to incorporate long-term bi-directional prediction to alleviate shortcomings of the previous algorithm. LP modelling-based measures, such as spectral flatness ration and pitch amplitude, have been used for sustained vowel data [35, 5] and for continuous speech data [36]. Recent investigations of disordered speech have used acoustic spectral and cepstral based measures that can be applied effectively to connected speech samples across a wide range of dysphonia severity [37]. It has been found that these techniques provide a better correlation with subjective ratings of dysphonia for both sustained vowel and connected speech, compared to traditional time-based acoustic methods [38, 39, 5, 40]. In particular, the Cepstral Spectral Index of Dysphonia (CSID)[1] has been put forth as an objective treatment outcomes

[1]CSID is a cepstral/spectral-based acoustic measure contained within the clinically-available Analysis of Dysphonia in Speech and Voice (ADSV) program.

measure of dysphonia severity [41]. The CSID is computed as the linear combination of means and standard deviations of cepstral peak prominence and the ratio of low frequency to high frequency energy, calculated across the sentence or the sustained vowel [41]. Recently, Peterson *et al.* [42] examined the validity of CSID in predicting the voice quality ratings for a variety of disorders. The results of their study indicated robustness of the CSID with respect to the dysphonia severity and some diagnostic categories. However, the study revealed as well some shortcomings of the CSID method. Although a correlation coefficient of $0.81 - 0.83$ between perceived (subjective) and CSID estimated ratings was obtained for the sustained vowel samples, only a correlation of $0.66 - 0.67$ was observed between perceived (subjective) and CSID estimated ratings when considering connected speech. Moreover, as can be observed in Fig. 1.6, the predictability of CSID is sensitive to the change across the diagnostic categories tested.

The traditional acoustical techniques for the analysis of continuous speech samples (including the CSID) are generally based on the segmentation procedure to identify voiced, unvoiced and silent speech periods. To avoid this segmentation, it is possible to consider *non-stationary* analysis techniques [36, 43, 44]. Umapathy *et al.* [36] used the Matching Pursuit Algorithm (MPA) with Gabor dictionary to extract several features such as: octaves, energy capture ratio, length and frequency ratio. These time-frequency approaches were also applied to predict the quality of tracheoesophageal (TE) speech samples [45]. TE speech contains much more noise due to the neck breathing and irregular vocal fold vibrations. Although the results in [45] showed a modest correlation of 0.63 between the predicted and perceived speech quality ratings, this correlation was significantly better compared to traditional acoustic measures

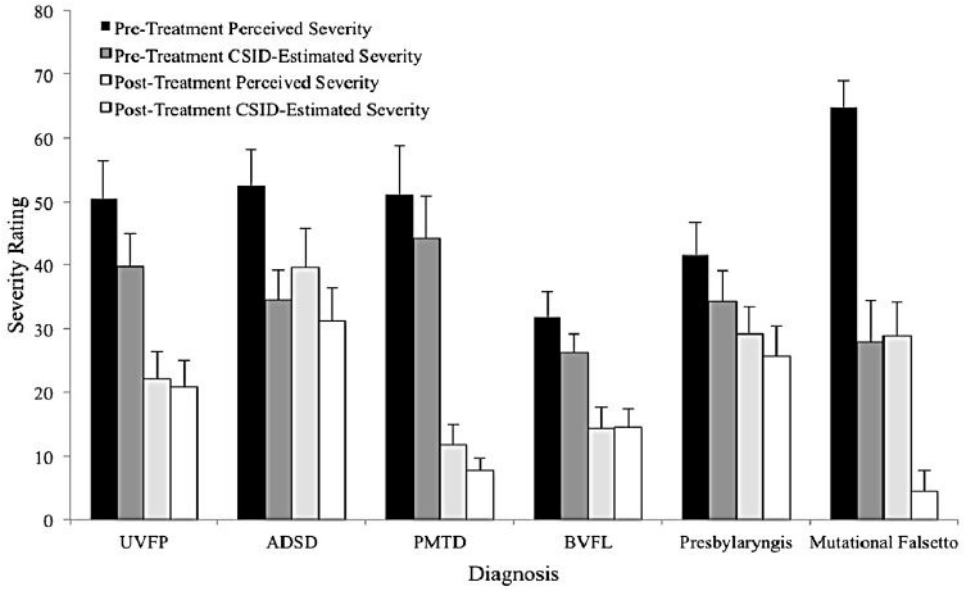

Figure 1.6: Perceived versus CSID-estimated severity ratings for connected speech by diagnostic category, unilateral vocal fold paralysis (UVFP), adductor spasmodic dysphonia (ADSD),primary muscle tension dysphonia (PMTD), benign vocal fold lesions (BVFLs) [42].

such as jitter, HNR parameters and linear prediction features.

1.6 Thesis Scope

Speech quality measurements are typically used to characterize the abnormality of voice. These measurements are utilized to assess the degree of voice disorder severity and to monitor the progress over the course of voice therapy. While subjective judgments of speech quality are the “gold standard" they are often time-consuming and resource-intensive. Objective speech quality measures – those that are computed from the raw acoustic data and correlate highly with subjective quality ratings – are therefore highly attractive, due to their efficiency and reliability. In addition, it is

also desirable that the objective measures are computed on connected natural speech samples to reflect their “real world" applicability. There is scope for improvement in the performance of current state-of-the-art objective measures used in clinical settings (e.g. the cepstral/spectral index of dysphonia) in terms of their correlation with subjective quality ratings of connected speech. This thesis therefore aims to develop and validate a novel and robust objective speech quality metric computed from connected speech samples. The proposed methodology incorporates aspects of speech production and perception models and the objective metric so derived is anticipated to outperform the current clinically available speech quality estimation tools.

1.7 Thesis Contributions

The thesis provides a novel non-intrusive estimation scheme for disorder voice that is based on an intrusive algorithm. Our idea consists of generating a reconstructed (i.e.,an approximation) of the voice/speech signal by running an adaptive time-frequency algorithm, Matching Pursuit (MP) [46], using a given number of iterations on the original disordered speech sample. The resulting reconstructed signal is then used as a reference signal for the intrusive algorithm, a process that results in the generation a quality score. The score is then used as our feature for disordered voice quality estimation. We consider the use of either the Perceptual Evaluation Speech Quality (PESQ) [47], which is the standard intrusive algorithm in telecommunication, or the Hearing Aid Speech Quality Index (HASQI) [48], used in the hearing aid industry, to perceptually compare the generated reference signal with the original degraded signal and extract a quality score. Both approaches are compared and studied using two

experimental databases of 24 and 35 speech samples obtained from patients who had undergone total laryngectomy and used tracheoesohageal (TE) voice. The results obtained show that our proposed estimation scheme perform significantly better than conventional acoustical measures used in voice quality research. This work has been published in [49].

As a second contribution, our goal is to propose acoustical features which are easily extracted (computationally simple) from a given speech signal. A novel low complexity algorithm to estimate the degree of severity of disordered TE speech is presented. The proposed algorithm uses features which are computed from 32- ms voiced frames of the speech signal. First, the voiced frames of the acoustical speech signals are extracted using the simple autocorrelation method [14] and the corresponding pitch estimation per voiced frame is obtained. A 18-th order LPC analysis, based on the Levinson-Durbin algorithm, is performed on each voiced frame of the speech. Then, we extract two groups of acoustical features: statistical features and vocal tract-based features. The group of statistical features consists of higher order statistics (central moments: mean, standard deviation, skewness and kurtosis) which are extracted from the LPC coefficients, cepstral coefficients and the LPC residual signal. The averages of each of these moments are computed along with the pitch average over all voiced frames yielding a total of 14 quality features. The second group of vocal tract-based features consists of predicting the model for the speech production system that generated the disordered speech at hand. This is done by calculating 16 cross sectional areas which are obtained from the LPC coefficients and 49 acoustical feature are extracted. A database of 35 TE speech samples is used to train and validate a linear regression model that combines all these features. Stepwise linear regression is first

used to prioritirize the most signficant features and then K-folds cross validation is used to reduce the number of features down to 20. The obtained set of features is used to train a model using support vector machines. The obtained model showed that the proposed speech quality estimation approach performs well with a correlation with subjective scores in the range between 0.81 and 0.86. This work has been published in [50].

As a third contribution, our goal was to develop a TE speech quality estimator that is *robust* across a number of different TE datasets collected at different conditions. We proposed a new approach for the estimation of the quality of TE speech using deep neural networks (DNNs). First, a synthetic dataset that consists of 2173 samples was used to train a DNN model that was shown to predict the TE voice quality. The synthetic dataset was formed by mixing 53 normal speech samples with modulated noise signals that had a similar envelope to the speech samples, at different speech-to-modulation noise ratios. A validated instrumental speech quality predictor was used to quantify the perceived quality of speech samples in this database, and these objective quality scores were used for training the DNN model. The synthetic dataset was divided into three subsets representing training, cross validation, and test datasets that contained respectively 70%, 15%, and 15% of the whole dataset. The DNN model was comprised of an input layer that accepted sixty relevant features extracted through filterbank and linear prediction analyses of the input speech signal, two hidden layers with 15 neurons each, and an output layer that produced the predicted speech quality score. The DNN trained on the synthetic dataset was subsequently applied to four different databases that contained speech samples collected from TE speakers. The DNN-estimated quality scores exhibited strong correlation with the

subjective ratings of the TE samples in the four databases, with correlation values reaching up to 0.8 and exceeding the performance of most existing TE speech quality estimators from the literature.

1.8 Thesis Outline

This thesis is organized as follows:

Chapter 2 provides the necessary background and preliminary material corresponding to the signal processing tools used throughout the paper.

Chapter 3 introduces a novel disordered speech quality estimation scheme using the Matching Pursuit algorithm. After an introduction, Section 3.2 recalls the full reference speech quality estimation method. We particularly dicuss the Perceptual Evaluation of Speech Quality (PESQ) and the Hearing Aid Speech Quality (HASQI) algorithms. In Section 3.3 the adaptive time-frequency matching pursuit algorithm is explained and discussed. Section 3.4 presents our approach to the disordered speech quality estimation that combines the Matching Pursuit algorithm with a full reference perceptual quality estimator. Finally, Section 3.5 presents some experimental results that validate the approach on TE speech databases.

Chapter 4 is devoted to speech quality feature extraction using acoustical analysis. Section 4.2 discussed the background on Linear Prediction analysis and different methods for regression and model fitting. In Section 4.3, we present the algorithm for pitch estimation and voiced/unvoiced frames extraction from a disordered speech signal. Then, in Section 4.4, we present different features which are extracted from a linear prediction analysis on the voiced frames of the speech signal. Section 4.5

presents different models which are fit to a TE speech database and shows a high correlation value with subjective scores.

Chapter 5 presents a novel speech quality estimation using deep neural networks (DNNs). Section 5.2 recalls the deep learning method. Section 5.3, presents the methodology that we followed to generate a DNN model capable of predicting the speech quality without requiring a reference signal. We present in Section 5.4 the obtained experimental results which showed the robustness and the high performance of the proposed DNN-based algorithm.

Chapter 6 summarizes the findings of this thesis and presents some possible future directions for research.

Chapter 2

Background and Preliminaries

2.1 Signal Processing Tools

In this section we present some signal processing tools used throughout the thesis.

2.1.1 Forward Stepwise Regression

Forward Stepwise Regression (FSR) is an approach of Stepwise Regression (SR) fitting model that uses an automatic procedure to choose the predictive variables. There exist other approaches for SR as backward and bidirectional elimination. The FSR is, typically, used when a large group of variables exists in order to provide a first screening of the candidate variables. It consists of the following steps:

- Start with no candidate variable in the model.
- By using a model fit criterion, the addition of each variable is tested.
- The variable whose inclusion gives the most statistically significant improvement

of the fit is chosen.

- Stop the process when none of the remaining variables are significant.

2.1.2 K-folds cross-validation

K-folds cross-validation is an approach that combines measures of fitness in prediction to derive a more accurate estimate of model prediction performance. Like other cross-validation approaches, K-folds cross-validation consists of dividing the sample data into complementary subsets, performing the analysis on one subset (called a training set) and validating the analysis on the other subset (called a validation set).
This technique can be performed in several steps

- The sample is randomly partitioned into K-equal sized subsamples.
- A single subsample is retained for validation data and $K - 1$ subsamples are used as training data.
- The procedure is repeated K-times, with each of K subsamples used precisely once as validation data.
- The average of K-result can be used as a single estimation.

One of the advantages of this method is that all observations are used for both training and validation and each observation is used for validation exactly once.

2.1.3 Mean Squared Error

The Mean Squared Error (MSE) is a risk function that assesses the quality of a predictor or an estimator. It measures the average of the squares of the error between

the observed values and the predicted ones. The MSE is always non-negative, and values closer to zero are better. Consider a sample of n data points, the MSE of a predictor that maps the observed values vector Y to the predicted values scalar $\hat{Y}$ is given as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \tag{2.1}$$

2.1.4 Linear Prediction

Linear Prediction (LP) is a very powerful tool for estimating the parameters of speech models. The speech production system can be modelled as input-filter-ouput model where the vocal tract acts as a time-varying filter. If one takes a small enough speech segment, it is resonable to assume that the vocal tract model is a linear time-invariant filter (constant transfer function). In linear prediction analysis, it is assumed that current speech samples are approximated by a linear combination of past speech samples which translates to an autoregressive model [51] in the signal processing community. The predictor coefficients are the weighting coefficients used in the linear combination of past speech samples and are derived by minimizing the sum of squared differences between the actual speech samples and the predicted ones. In this thesis we will focus on *forward linear prediction* technique where the term forward will be ommited for simplicity.

Suppose we wish to predict the value of the speech sample $x(n)$ using a linear combination of p most recent past samples. The estimate has the following form:

$$\hat{x}(n) = \sum_{i=1}^{p} a_{i,p}x(n-i). \tag{2.2}$$

The integrer p is called the *prediction order* and the coefficients $a_{i,p}$ are referred to as the p-th order linear prediction coefficients (see Figure 2.1 for a block diagram illustration of the LP). If the order of the LP analysis is understood from the context, the coefficients $a_{i,p}$ are called linear prediction coefficients for short. The estimation error, also called the *residual signal*, is defined as follows:

$$e_p(n) := x(n) - \hat{x}(n). \tag{2.3}$$

The linear prediction coefficients are computed by minimizing the following *mean square error:*

$$E_P = \sum_{n=0}^{+\infty} e_p^2(n). \tag{2.4}$$

Usually, linear prediction analysis is executed on short duration speech signals (called frames) which are obtained by *windowing* the data, *i.e.* multiplying the speech signal by a Hamming or similar time window. If we assume the speech segment to be multiplied by a window of length N which is zero outside the interval $0 \leq n \leq N-1$, then the mean square error reads:

$$E_P = \sum_{n=0}^{N-1} e_p^2(n). \tag{2.5}$$

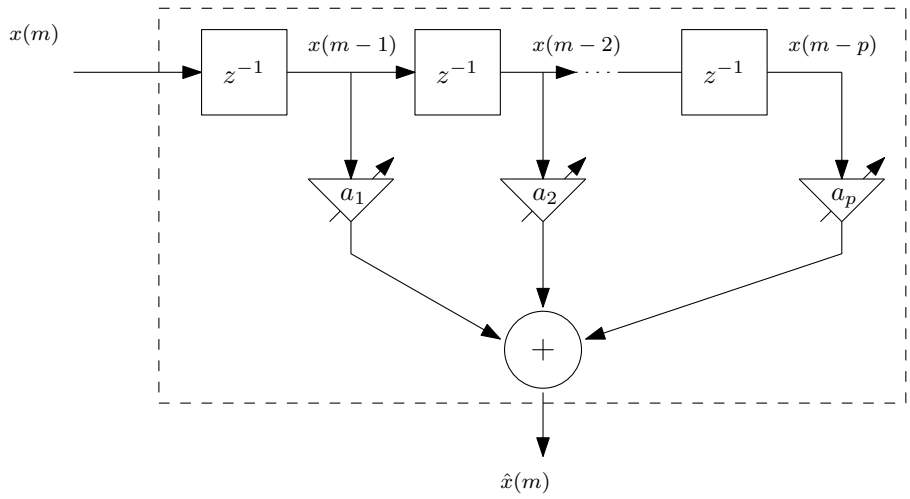

Figure 2.1: An illustration of the linear prediction process.

2.1.5 Pearson correlation

Pearson's correlation, also known as the Pearson Product-Moment Correlation, is defined as the ratio of the covariance of two variables X and Y representing a set of numerical data, normalized to the square root of their variances. It has a value between -1 and $+1$, where -1 is total negative linear correlation, 0 is no linear correlation, and -1 is total positive linear correlation. Consider a set of two-dimensional data points $[x_1, \cdots, x_N]$ and $[y_1, \cdots, y_N]$, the covariance between two variables X and Y is defined as

$$cov(X, Y) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}), \tag{2.6}$$

with $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$, $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$. Thus the Pearson correlation of two variables X and Y is defined as the following

$$\tau_{xy} = \frac{cov(X, Y)}{\sqrt{cov(X, X)cov(Y, Y)}} \tag{2.7}$$

2.1.6 Spearman rank correlation

The Spearman rank correlation, denoted ρ_{spear}, which is computed in a manner similar to ρ but with the original data values replaced by their ranks.

2.1.7 Sigmoidal mapping function

Sigmoidal mapping function was used and once the objective values were mapped, a new Pearson correlation (termed ρ_{sig}) was computed and used as the third performance criteria [52]. The sigmoid mapping is given by:

$$Y = \frac{\alpha_0}{1 + \exp(-(\alpha_1 X - \alpha_2))} \tag{2.8}$$

where α_0, α_1 and α_2 are the fitting parameters, X represents the predicted quality score, and Y the mapped predicted quality score. Lastly, the root square mean error, denoted RMSE, between the subjective and objective quality scores was used as our fourth performance criteria.

2.1.8 Higher-Order-Statistics

We refer by the term higher-order statistics (HOS) to functions which use the third or higher power of a sample. Examples of HOS are the third and higher moments, as used in the skewness and kurtosis, whereas the first and second moments, as used in the arithmetic mean (first), and variance (second) are examples of lower-order statistics. Despite the fact that HOS is significantly less robust than lower-order statistics due to the higher powers, the HOS has several uses, especially, in the estimation of shape parameters, such as skewness and kurtosis. The skewness is

a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. Its value can be positive or negative, or undefined. In case of a unimodal distribution, positive skew commonly indicates that the tail is on the right side of the distribution, and negative skew indicates that the tail is on the left. In a similar way, kurtosis is a descriptor of the shape of a probability distribution. The interpretation of the kurtosis depends on the its used particular measure. Given a real vector $x = \{x_k\}_{1 \leq k \leq K}$, where K is the dimension of x, we define its HOS (skewness and kurtosis) as follows:

$$\gamma_x = \frac{\frac{1}{K}\sum_{k=1}^{K}(x_k - \mu_x)^3}{\sigma_x^3},$$
$$\kappa_x = \frac{\frac{1}{K}\sum_{k=1}^{K}(x_k - \mu_x)^4}{\sigma_x^4}.$$

with μ_x and σ_x define the mean and standard deviation, respectively, and can be expressed as the following

$$\mu_x = \frac{1}{K}\sum_{k=1}^{K} x_k,$$
$$\sigma_x = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(x_k - \mu_x)^2},$$

2.1.9 The ITU-T P.862 standard

The Perceptual Evaluation of Speech Quality (PESQ) is an intrusive objective method for speech quality estimation. It has been widely used in telephony [47]. Different

versions of the algorithm have been standardized since the first recommendation ITU-T P.861 (PSQM) in 1997. The wideband version is the ITU-T P.862 standard. The general schematic of the algorithm is given in Figure 2.2.

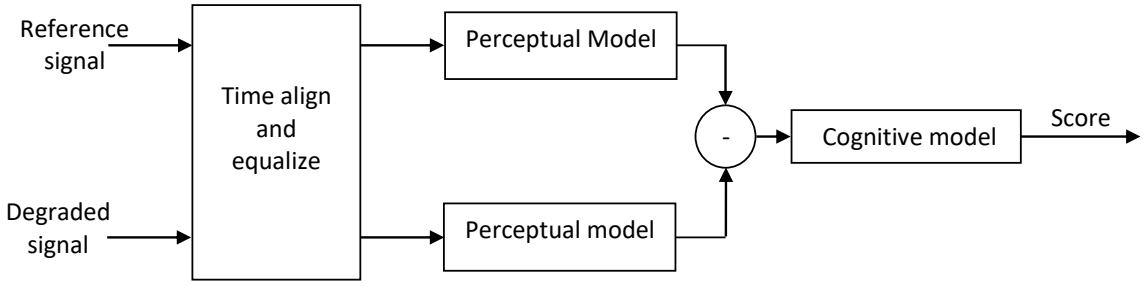

Figure 2.2: Block diagram of the Perceptual Speech Quality Evaluation (PESQ) computational procedure.

The PESQ algorithm is quite complex because it is composed of many steps. A high-level description of the important stages is as follows:

- **Preprocessing**: the first stage of the PESQ algorithm is the preprocessing, alignment and equalization. Here, both the degraded and original (reference) signals are aligned to the same constant power level and the same time (support)[1].

  - The level alignment algorithm proceeds as follows. First, filtered versions of the original and degraded signal are computed. Next, an average value of the squared filtered original speech samples and filtered degraded speech samples are computed. Finally, different gains are calculated and applied to align both the original and degraded speech signal to a constant target

[1]PESQ assumes that the subjective listening level is a constant 79 dB SPL at the ear reference point.

level resulting in the scaled versions of these signals.

  - The time alignment routine provides time delay values to the perceptual model to allow corresponding signal parts of the original and degraded files to be compared. This alignment process takes a number of steps. First, an envelope-based delay estimation is applied on the entire original and degraded signals. Next, the original signal is divided into a number of subsections known as utterances. Then, an envelope-based delay estimation is applied on utterances and the delay to nearest sample is identified using fine correlation/histogram-based algorithm. Finally, the utterances are split and the time intervals are realigned to search for delay changes during speech.

- **Psychoacoustic domain:** At this stage, both the degraded and original signals are transformed to the psychoacoustic domain using the perceptual model. This transformation is necessary to calculate the *distance* (comparison) between the two signals. First, the acoustic signals are mapped to the time-frequency domain using a short-term fast Fourier transformation (FFT) with a Hann window of size 32 ms with a 50% frame overlap. Next, the frequency axis is converted to the Bark scale [53]. The Bark scale can be interpreted as the perceived frequency scale within the human hearing system. In fact, the perceived change in frequency is not linear to the actual frequency change of the ear. Lower frequencies tend to have a finer resolution (in terms of the perception of sounds) than higher frequencies. After converting the frequency axis to the Bark scale, the intensity axis is warped to the *loudness* scale (Sone) using Zwicker's law [54].

- **Comparison:** once the two signals (degraded and original) are mapped to the psychoacoustic domain, they can be subtracted to obtain the disturbance density for each time-frequency cell. The disturbance densities are then conditioned accordingly to take into account the fact that small differences (distortions) are inaudible in the presence of loud signals (masking effect). The disturbance values are then averaged to yield the final PESQ score which ranges between -0.5 to 4.5.

2.1.10 The Hearing Aid Speech Quality Index (HASQI)

Kates and Arehart [48] have developed an objective measure (index) of speech quality estimation, named the Hearing Aid Speech Quality Index (HASQI). Although HASQI was originally developed to evaluate the effect of distortions introduced by hearing aids on the quality of the speech perceived by a hearing impaired listener, it has also been reported have valid applications on the performance of listeners without hearing loss, even for perception of non-speech sounds [55, 56]. The authors also have recently proposed a new version for the HASQI, [57] which includes computational efficiency and accuracy indices for both normal-hearing (NH) and hearing-impaired (HI) listeners.

The index mainly captures the effects of four components, divided into two indices: noise and nonlinear distortion (called the nonlinear index) and linear filtering and spectral changes (called the linear index), see Figure 2.3.

The nonlinear index computes the time-frequency representations for both the original and degraded speech signals using a basic cochlear model. It combines the cepstral correlation with a vibration correlation term.

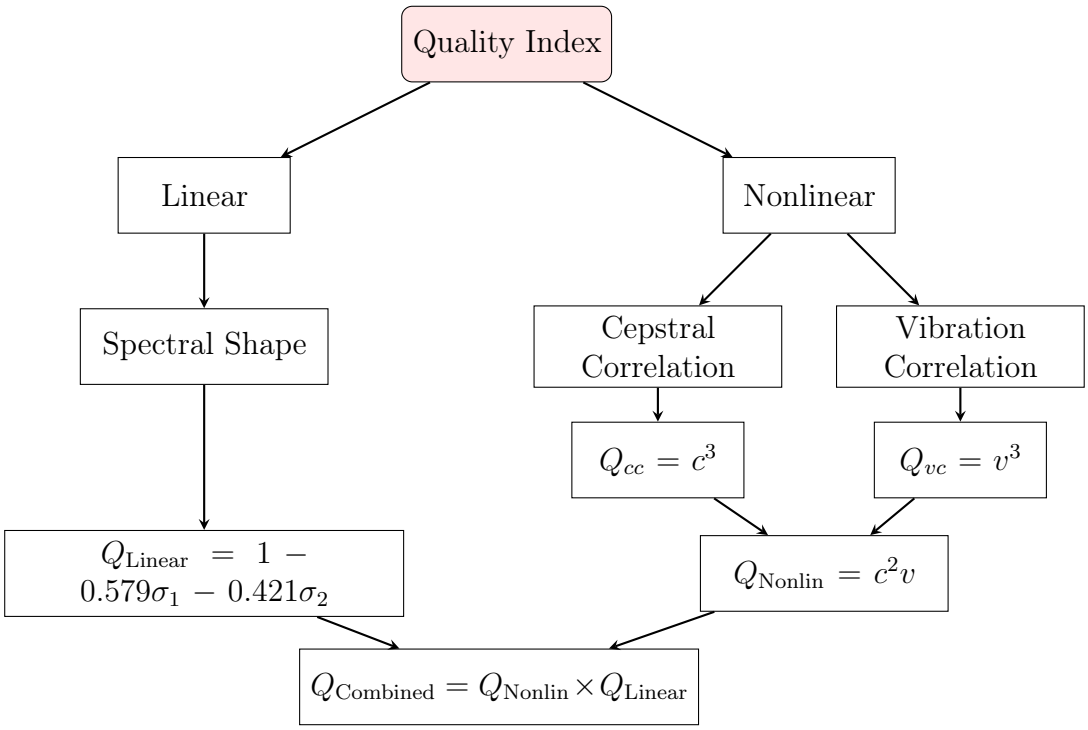

Figure 2.3: Diffrent types of indices used in the HASQI process.

- **The cepstral correlation** is computed by taking the cross-correlation of the cepstral sequences for the reference and degraded signals. The nonlinear index using just the cepstral correlation is a MMSE third-order regression fit to the combined NH and HI subject ratings for the noise and distortion stimuli and is given by

$$Q_{cc} = c^3, \tag{2.9}$$

  where c is the cepstral correlation.

- **The vibration correlation** is computed by taking the cross-correlation of the segmented basilar membrane (BM) vibration signal for the reference and degraded signal. It measures changes of the signal over time (temporal fine

structure) while ignoring any long-term spectral change. The nonlinear index using just the vibration correlation is a MMSE third-order regression fit to the combined NH and HI subject ratings for the noise and distortion stimuli and is given by:

$$Q_{vc} = v^3, \tag{2.10}$$

where v is vibration correlation.

- **Cepstral correlation $\times$ vibration correlation** Kates and Arehart [48] found that the product of the square of the cepstral correlation index times the BM vibration index is the most accurate combination of the first-order and second-order terms and cross products, and is given by

$$Q_{\text{Nonlin}} = c^2 v, \tag{2.11}$$

In the case of the linear index, the time-frequency representations are averaged across time and the index provides how large the differences are between the long-term average spectra (LTAS) of the test signal and the reference signal while ignoring the short-term differences in signal modulation and temporal fine structure. This linear index is a MMSE linear regression that is fit to the combined NH and HI listener ratings for the linear filtered stimuli, and is given by

$$Q_{\text{Linear}} = 1 - 0.579\sigma_1 - 0.421\sigma_2, \tag{2.12}$$

where σ_1 is the standard deviation of the differences in the spectral shape and σ_2 is the standard deviation of the differences in the spectral slope. The HASQI index is,

therefore, a multiplicative combination of the nonlinear and linear indices

$$Q_{\text{Combined}} = Q_{\text{Nonlin}} \times Q_{\text{Linear}}, \tag{2.13}$$

and is limited to lie between 0 and 1, with 0 indicating the poorest voive/speech quality predicted by the algorithm and 1 indicating perfect quality score.

Remark 1. *It has been shown that the multiplicative index is monotonic in its two constituent components and has no free parameters. Moreover, the multiplicative index is, in the absence of noise and distortion, identically the linear index, and in the absence of linear filtering is identically the nonlinear index.*

2.2 The ITU-T P.563 Standard

In this research, we propose to integrate speech production-based features with speech perception-based features into one single voice quality estimation model. Our approach is built on the non-intrusive quality measure standardized by the International Telecommunication Union (ITU) for telephony-related applications (ITU-T, P.563, [44]).

A schematic of the main functional blocks of the ITU-T P.563 standard is given in Fig. 2.4. As shown in Fig. 2.4, the speech sample under test is first passed through a preprocessing stage wherein level normalization, intermediate reference system (IRS) filtering, and voiced/unvoiced segmentation operations are performed. The processing then branches off to three parallel algorithm blocks which extract several features from the preprocessed speech signal that are later combined to produce the objective speech quality index.

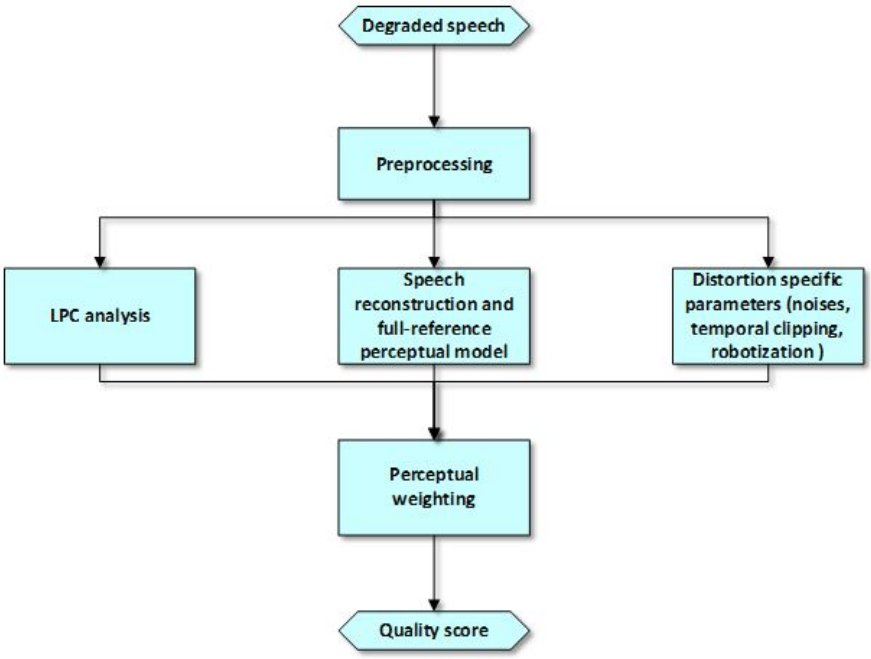

Figure 2.4: ITU-T P.563 schematic diagram.

The first of these three blocks employs LP analysis to model the variations in the voice production system (vocal tract). In particular, the voiced sections of the speech are modelled as a series of tubes with different lengths and cross-sections varying over time [58]. As shown in Fig 2.5, the first stage of the modelling consists of pitch extraction using the hybrid temporal/spectral method proposed by Gray *et al.* [59]. In the second stage, and for each pitch cycle, the vocal tract is modelled as eight tube sections whose section areas are calculated using the LP analysis. These eight section areas are averaged to model the cavity articulators: rear, middle and front cavities. It is worth pointing out that the vocal tract parameters are only calculated for the voiced part of the speech signal. In addition, higher order statistics (*viz.* skewness and kurtosis) of the LP coefficients and their cepstra are computed. This set of statistical measures, the vocal tract quality parameters, and the tracked section size

changes, are combined with others in the mapping module (perceptional weighting) of the P.563 to estimate the speech quality.

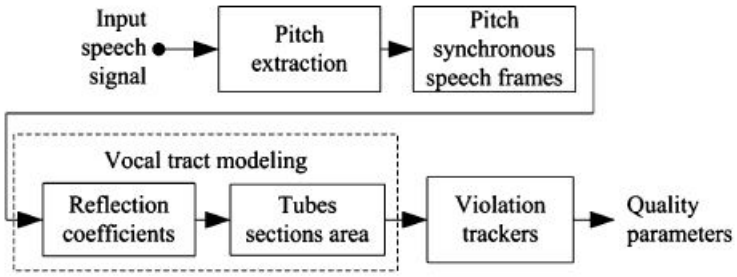

Figure 2.5: Pitch synchronous vocal tract model and LPC analysis [44]

The second main P.563 algorithm block encompasses LP-based reconstruction of a quasi-reference signal which, along with the original distorted signal, is given to an intrusive perceptual model as shown in Fig 2.6. The quasi-reference signal is generated by extracting LP coefficients from the input speech signal on a frame-by-frame basis, modifying these LP coefficients to conform to natural vocal tract constraints, and synthesizing the speech frames using the modified LP coefficients. The perceptual speech quality model incorporates a computational model for human audition and quantifies the perceptually-relevant differences between the input speech and the quasi-reference speech samples. The speech perception modeling part of the P.563 standard is based on another ITU-T standard, the ITU-T P.862 [47].

Finally, the third block generates a number of distortion-specific parameters such as temporal and amplitude clipping, interruptions and mutes, estimated low SNR for background noise, segmental noise, robotisation, and unnatural male and female speech. The final quality score is estimated by a linear, perceptually-weighted, combination of all these parameters [44]. Different weighting coefficients are chosen for

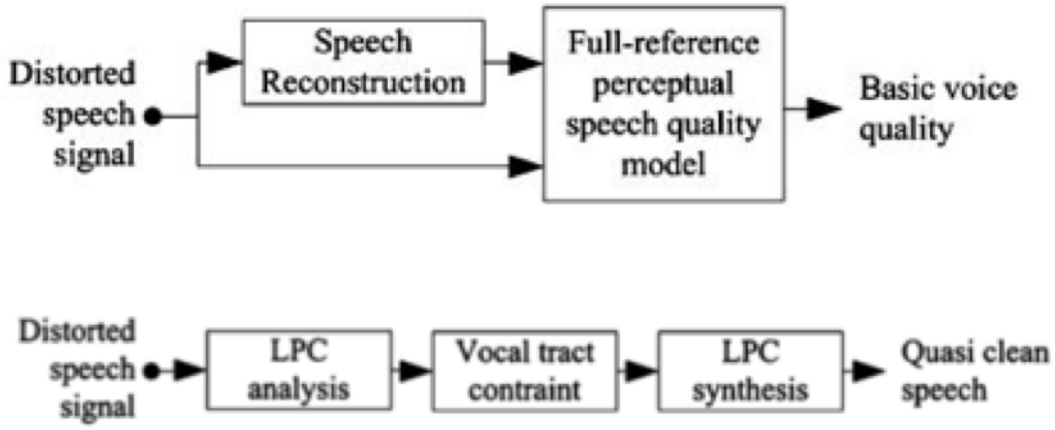

Figure 2.6: Full-reference quality assessment model in P.563 [44].

different classes of signal degradations to improve the speech quality estimation accuracy.

2.3 TE speech databases

In this section we describe the TE speech datasets used throughout the thesis. These datasets are collected at different times, different conditions and contain different number of TE samples. These datasets were collected and evaluated by researchers at the School of Communication Sciences and Disorders at Western University after obtaining ethics approval from the University's health sciences research ethics board.

The speech samples were recorded from adult males between the ages of 45-65 years. All had undergone total laryngectomy and TE puncture voice restoration and all were at least one-year postsurgery at the time of their participation. All recordings were gathered in a sound-treated environment at a sampling rate of 44.1kHz with 16-bit quantization. For datasets D1, D3 and D4, the second sentence of a standard

reading (The Rainbow Passage), "The rainbow is a division of white light into many beautiful colors" was extracted from the full recording from all speakers and used for acoustic and perceptual measurements. For the dataset D2, on the other hand, the first sentence of the Rainbow Passage, “When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow" was extracted. The statistics of these reference datasets are given in Table 2.1. Speech quality is a multi-dimensional perceptual phenomenon that encompasses attributes such as clarity, pleasantness, naturalness, etc [60].

Dataset	Measure (scale)	Samples	Minimum	Maximum	Mean	Standard Deviation
D1	severity (0-100)	24	19.6	90.85	54.77	22.41
D2	quality (0-100)	19	34.83	77.99	59.15	12.29
D3	severity (0-100)	20	14.51	88.186	47.96	20.33
D4	quality (1-10)	35	1.74	7.53	4.97	1.75

Table 2.1: Statistics of the subjective scores for the three reference TE speech datasets.

For the auditory-perceptual phase of the study, the TE speaker samples were played back to a group of naive listeners who had no prior exposure to TE speech. The signals were played back in a randomized order and the listeners were instructed to rate the overall perceived severity/quality on a visual analog scale. The average of listener ratings were then used as the final subjective score.

2.4 Conclusion

We presented some of the important signal processing and statistical analyses tools that are needed throughout the thesis. We also presented the standard ITU-T P.563 that is used as a non-intrusive algorithm for speech quality estimation in telecommu-

nication. Our developed algorithms in the next two chapters will be inspired from some of the building blocks of the ITU-T P.563 but tuned, optimized and tailored for the clinical applications related to TE speech quality estimation.

Chapter 3

Disordered speech quality estimation using the matching pursuit algorithm

3.1 Introduction

This chapter propose a novel non-intrusive estimation scheme for disordered speech quality that is based on the standard intrusive PESQ [47] and HASQI [48] algorithms. Our idea consists of generating a reconstructed (i.e., an approximation) of the voice/speech signal by running a matching pursuit algorithm [46], using a given number of iterations on the original degraded sample. The resulting reconstructed signal is then used as a reference signal for the intrusive algorithm, a process that results in the generation of a quality score. The score is then used as our feature for disordered voice quality estimation. Our approach is tested on the TE speech datasets described in Section 2.3. High correlation values (compared to the state-of-the-art algorithms) are obtained for two datasets while moderate correlations are obtained

for the other two datasets. The results presented in this chapter are based on our work in [50].

3.2 Quasi-Reference Perceptual Speech Quality Estimation

In this section we present the essential ingredients of the proposed methodology for TE speech quality estimation based on the matching pursuit algorithm. The overall idea consists of generating a quasi-clean reference signal extracted from the degraded speech signal using the matching pursuit algorithm that will be described in Section 3.2.1. The extracted quasi-reference signal is then fed (along with the original degraded signal) to the state-of-the-art intrusive speech quality evaluation algorihtms PESQ and HASQI described in Sections 2.1.9 and 2.1.10, respectively.

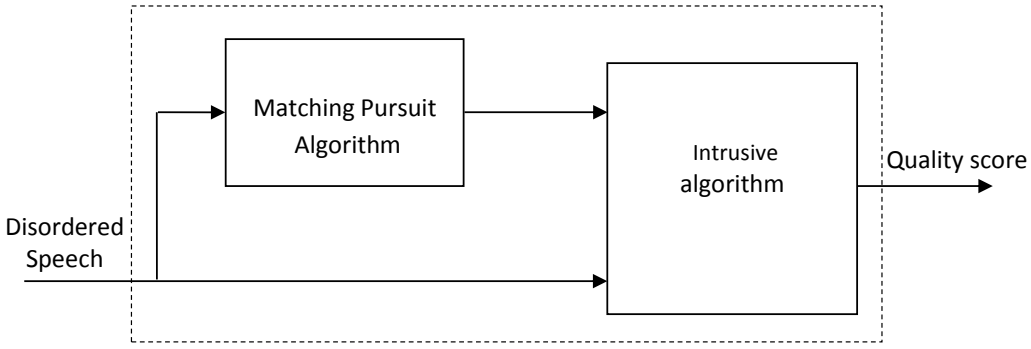

Figure 3.1: Schematic of the proposed quasi-reference intrusive speech quality estimation algorithm .

3.2.1 Matching Pursuit Algorithm

The matching pursuit algorithm (MPA) provides a way to expand, or represent, a signal in terms of any time-limited functions, or atoms, called dictionary [46]. The time-frequency atoms are a family of functions that are well localized in both time and frequency. The matching between the input and dictionary elements are used to select elements of the representation adaptively. The signal decomposition is based completely on the particular dictionary and the matching criterion. The best element is chosen at each step. After a number of decomposition, the original signal, denoted $x(t)$, can be represented to some arbitrary resolution by a series of expansion coefficients. The approximated signal reconstructed, at the $n-$iteration, from these expansion coefficients is given by

$$\hat{x}_n(t) = \sum_{i=0}^{n-1} \alpha_i h_{\gamma i}(t), \tag{3.1}$$

where x_n is the input signal, n is the number of iterations, α_i represents the expansion coefficients and $h_{\gamma i}$ is the time-frequency atoms.

For instance, consider the *Gabor* dictionary [61]

$$D = \left\{ h_\gamma(t) \triangleq \frac{1}{\sqrt{a}} h\left(\frac{t-b}{a}\right) e^{j2\pi ft} \right\}, \tag{3.2}$$

where $h(t) = \pi^{-\frac{1}{4}} e^{-\frac{t^2}{2}}$ represents the Gaussian window (Figure 3.2) and γ is the index set $\gamma = \{a, b, f\} = \{\text{scale,translation,frequency}\} \in R^+ \times R^2$. A Gabor atom can be seen, simply, as an atom that uses the Gaussian function. The family of time-frequency atoms can be generated by scaling, translating and modulating a

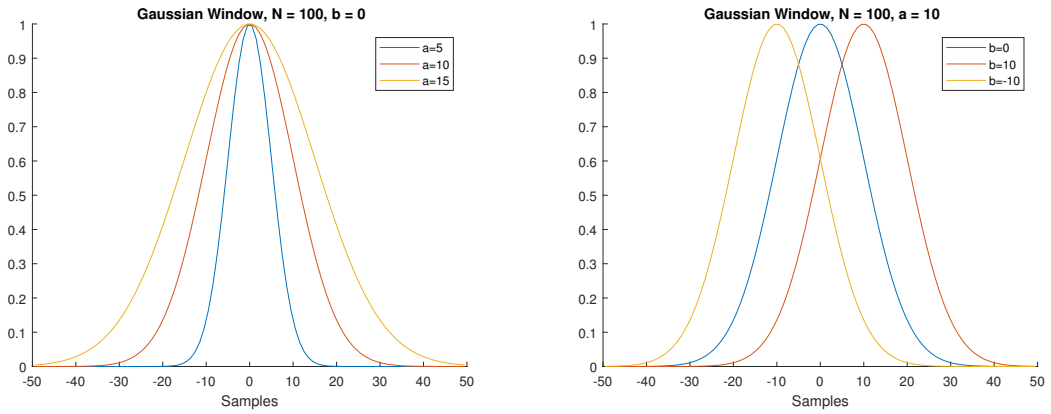

Figure 3.2: Illustration of the Gaussian window with different values of a and b.

single window function. The scale, location and frequency can all be independently adjusted, thus, providing large flexibility compared to the wavelet transform analysis. Selection criterion used to select dictionary elements on any given iteration is based on the magnitude of the inner product of the input $x(t)$ and the dictionary elements,

$$< x(t), h_\gamma(t) >= \frac{1}{\sqrt{a}} \int_0^\infty x(t) h\left(\frac{t-b}{a}\right) e^{-j2\pi ft} dt. \tag{3.3}$$

For each iteration, the atom $h_\gamma(t)$ maximizing the above inner product is retained. The residual approximation error at the i-th iteration is given recursively by $R^{i+1}x(t) = R^i x(t) - < x(t), h_\gamma(t) > h_\gamma(t)$. The algorithm further approximates the residue $R^{i+1}x(t)$ by selecting another best atom $h_{\gamma'}$ from the dictionary. This process is repeated for a given number of iterations n to produce the following signal decompo-

sition

$$x(t) = \hat{x}_n(t) + R^n x(t), \quad R^0 x(t) = x(t) \tag{3.4}$$

$$\hat{x}_n(t) = \sum_{i=0}^{n-1} \langle R^i x(t), h_{\gamma_i}(t) \rangle h_{\gamma_i}(t) \tag{3.5}$$

An illustration example of the procedure is shown in Figure 3.3 where we can see the original signal with (at top left) the right top first time-frequency atom used to decompose it. The first residual signal is represented in the bottom plot.

3.2.1.1 Energy Capture Ratio (ECR) feature

The Energy Capture Ratio (ECR) is a simple speech quality estimation feature that was proposed in [43] based on the matching pursuit algorithm. It is defined at the n-th iteration by

$$\text{ECR} = \frac{\sum_{i=0}^{n-1} \alpha_i^2}{E}, \tag{3.6}$$

where E is the energy of the original signal $x(t)$; assumed finite, and $\alpha_i = \langle R^i x(t), h_{\gamma_i}(t) \rangle$. Theoretically, if we run the MPA for an infinite number of iterations and also using a large enough dictionary of atoms then one would have

$$\lim_{n \to \infty} \sum_{i=0}^{n-1} \alpha_i^2 = E, \tag{3.7}$$

or, equivalently, the signal $x(t)$ has been fully reconstructed using *infinite* linear combinations of atoms. This feature will be also used and compared against the proposed speech quality features of this chapter.

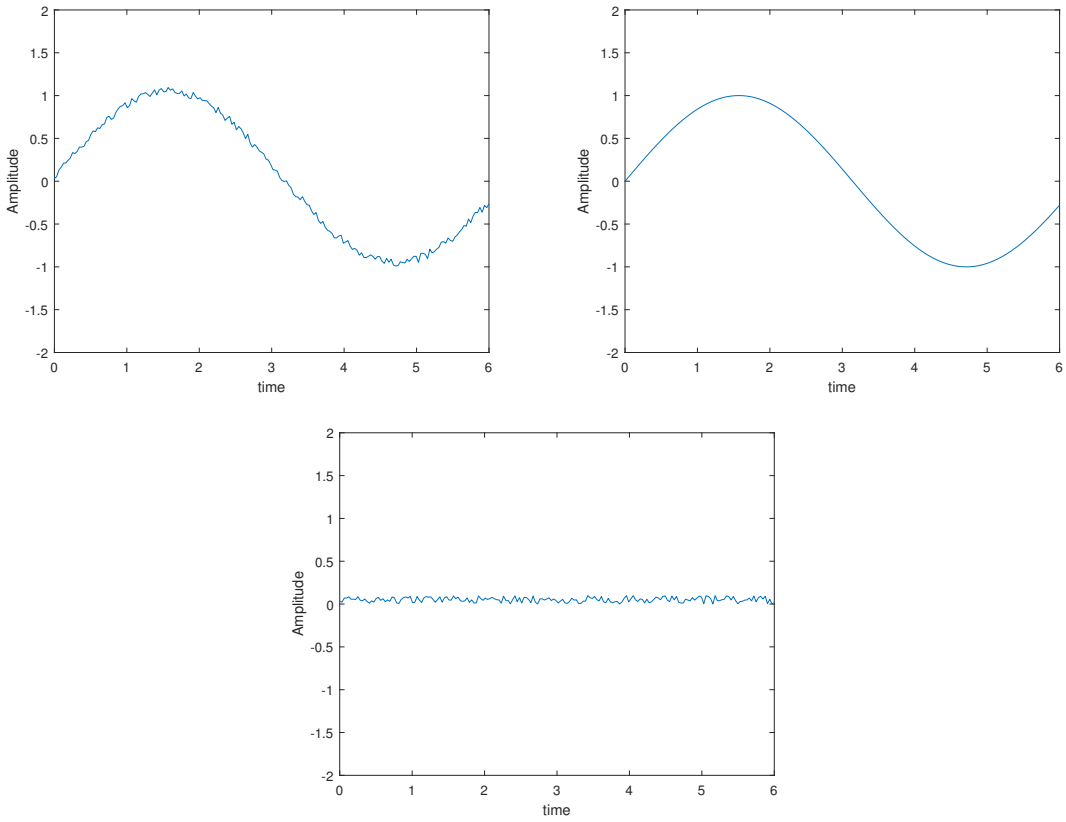

Figure 3.3: Illustration of the matching pursuit algorithm. a) Input (original) signal b) The used time-frequency atom c) Residual signal after removal of energy represented by the time-frequency atom.)

3.2.2 Proposed speech quality estimation features

In this section, we discuss the newly proposed speech quality estimation features derived using the MPA algorithm, the perceptual speech quality estimator (PESQ) and the hearing aid speech quality index (HASQI) discussed in the previous subsections.

Our derived features are based on the following observation: the closeness of the approximation signal $\hat{x}_n(t)$ to the original signal $x(t)$ is dependent on the number of iterations n – more atoms are required to model the transients and nonstationary

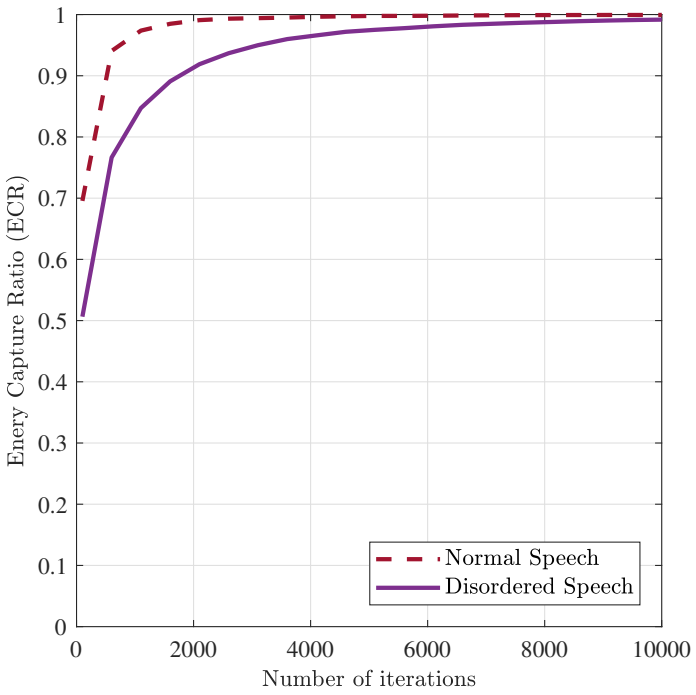

Figure 3.4: Energy capture ratio for normal and disordered speech samples.

portions of the input signal. Since disordered voice/speech signals are generally characterized by greater degree of nonstationarity, a greater number of atoms are required for reconstruction. This phenomenon is displayed in Figure 3.4, where the closeness of reconstruction, measured by the Energy Capture Ratio (ECR) [36], is plotted against the number of atoms for a normal clean speech signal and a “white noise" disordered voice sample. It is clear that normal speech can be approximated to 99% using 1100 atoms, while the disordered sample requires more than 8300 atoms for the same level of reconstruction. Stated differently, the reconstructed signal using N atoms will be closer to the original input if it is normal speech, and farther if it is abnormal. Based on the above discussion, we propose the following speech quality features:

3.2.2.1 MPA-PESQ feature

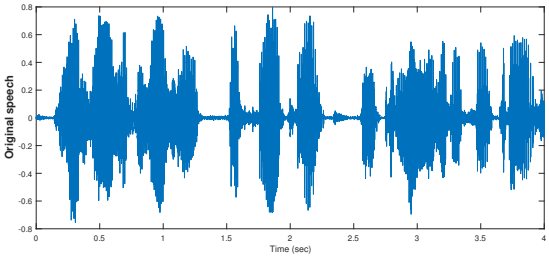

Figure 3.5: An original speech sample from the first Database D1.

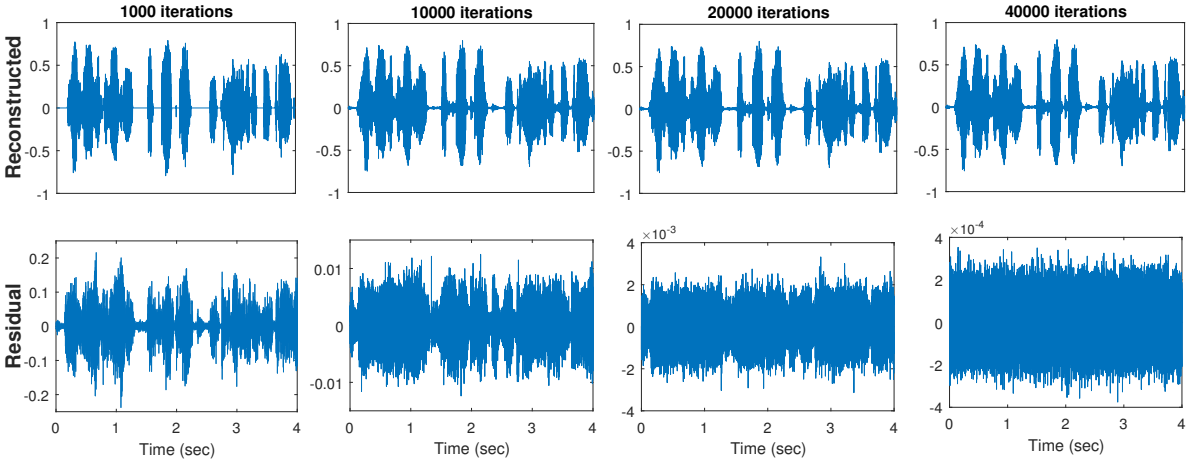

Figure 3.6: Reconstruction of the speech data sample in Figure 3.5 using the MPA at different iterations.

The reconstructed signal, using the MPA, is used as the quasi-reference input to the speech perception model of the ITU-T P.862 standard, which will return a high quality score for normal voice samples (due to the "closeness" of the original and quasi-reference inputs) and a lower quality score for disordered voices (due to the increased "distance" between the original and quasi-reference inputs). At this stage the issue that remains to be investigated is the search for the most appropriate or *best* number

of iterations to be used on a degraded speech signal to generate a “good" or suitable reference signal from the reconstruction algorithm. As discussed above, as the number of iterations of the MPA increases, the difference between the reconstructed and the original voice samples tends to decrease. Figure 3.6 shows the difference between the original degraded signal and the reconstructed signal when running the MPA using an increasing number of iterations $(1000, 10000, 20000$ and $40000)$. As expected, as the number of iterations increases, the difference between the reconstructed and the original samples tends to decrease. Of course, in our application, it is useless to *fully* reconstruct the degraded voice signal using a large number of iterations. Also, a very low number of iterations would be insufficient as we will end up with *non-modelled* speech components (see the entry on the far left of Figure (3.6)). Our idea consists in finding an *optimal* number of iterations which allow us to reconstruct the coherent components of the degraded voice signal while leaving the non-coherent noisy-like components in the residual signal. When the number of atoms is 20000 and more, we can see that the residual signal does not contain any more speech-like components and only white noise-like components are present. Of course, going up to 40000 atoms we see that the original signal is almost reconstructed and only white-noise is present in the residual. This will not serve us in our application to highly disordered voice samples, as we are looking to avoid modelling all the distortions which are likely to be at high frequencies. As a first intuition, the best number of iterations would be something between 10000 and 20000. Later on, we will increase the number of iterations in an effort to provide the best possible correlation for our database. Thus our voice quality estimation method consists of two stages:

- Generate a quasi-reference signal using the MPA from the given disordered voice

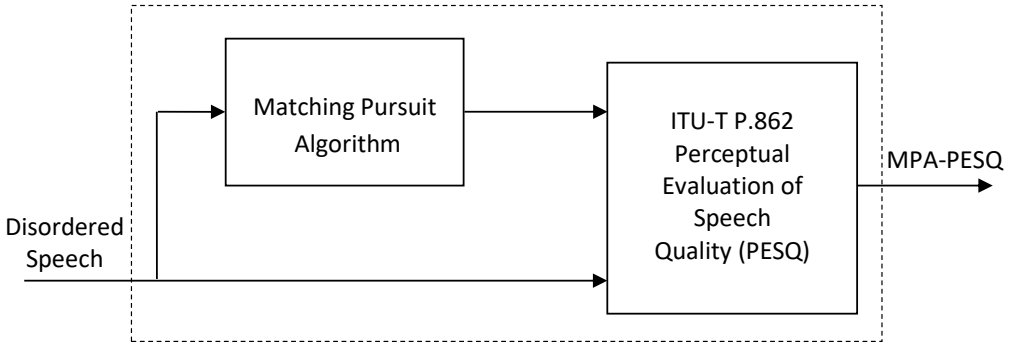

Figure 3.7: Schematic of the MPA-PESQ quality feature.

signal.

- Use the obtained quasi-reference signal along with the original disordered voice signal, in combination with a perceptual speech quality assessment algorithm such as the PESQ.

3.2.2.2 MPA-HASQI feature

Similarly to the MPA-PESQ feature, the reconstructed voice signal, using the MPA, can be used as the quasi-reference input to the hearing-aid speech quality index HASQI. Doing so will also return a high quality score for normal voice and a lower quality score for a disordered sample. Note that the original HASQI scale ranges from between 0 and 1:, in our experiments, we found that the interval of the output for the MPA-HASQI feature depends strongly on the number of iterations used in the MPA.

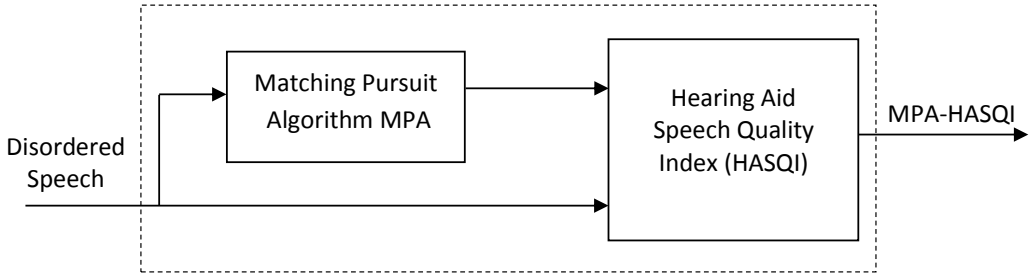

Figure 3.8: Schematic of the MPA-HASQI quality feature.

3.2.2.3 Residual Rate Feature (RRF)

Here we present a third speech quality estimation feature that is extracted from the rate of decrease of the ECR feature. In fact, one can write

$$\mathrm{ECR} = 1 - \underbrace{\frac{\sum_{i=n}^{\infty} \alpha_i^2}{E}}_{r(n)}, \tag{3.8}$$

where the residual energy ratio $r(n)$ clearly satisfies the two properties

$$r(0) = 1, \tag{3.9}$$

$$\lim_{n\to\infty} r(n) = \lim_{n\to\infty} \sum_{i=n}^{\infty} \alpha_i^2 = 0. \tag{3.10}$$

As argued in [36], any portion of a good quality speech signal will be modeled in fewer iterations compared to a poor quality voice sample. Therefore, it is expected that the rate of decrease of the residual energy ratio function against the number of iterations for a good quality voice signal will be larger than its counterpart for a lesser quality sample.

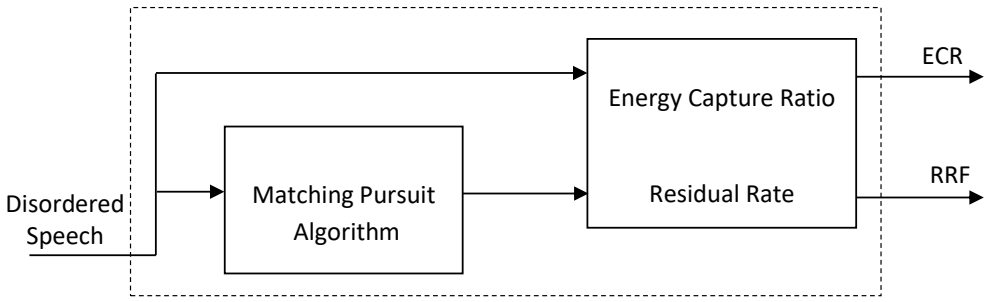

Figure 3.9: Derived matching pursuit features ECR and RRF based on the energy capture of the original and reconstructed signal.

As an approximation, we assume that the residual energy ratio function is modeled by an exponential function decaying at a constant rate, thus

$$r(n) \sim \hat{r}(n) := e^{-kn}, \tag{3.11}$$

for some $k > 0$. The constant rate k will be characterizing the voice signal, *i.e.,* different speech signals will have different rates. The greater the value of k, the better the voice quality represented. Therefore, the rate k can be taken as a new quality feature

$$\text{RRF} = k. \tag{3.12}$$

Now, we compute the value of k for a given speech signal. Using each number of iterations $n = n_{\min} + j \cdot h,\ j \in \{0, 1, \cdots, N\}$ where N and h are some positive integer numbers, we execute the MPA and compute the residual energy ratio $r(n)$. Then,

using linear regression (least-mean square method) we find the best scalar $k > 0$

$$\begin{bmatrix} \log(r(n_{\min})) \\ \log(r(n_{\min}+h)) \\ \vdots \\ \log(r(n_{\min}+(N-1)h)) \\ \log(r(n_{\max})) \end{bmatrix} \simeq -k \begin{bmatrix} n_{\min} \\ n_{\min}+h \\ \vdots \\ n_{\min}+(N-1)h \\ n_{\max} \end{bmatrix} \tag{3.13}$$

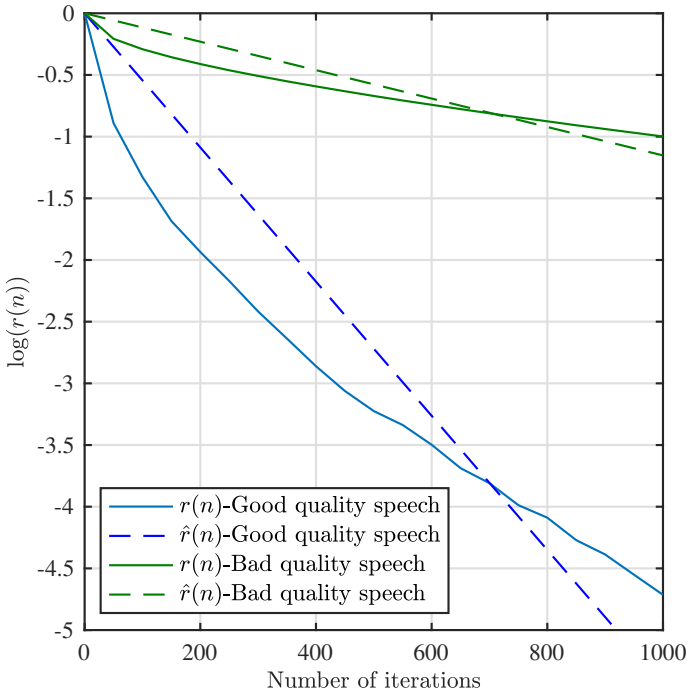

Figure 3.10: Logarithm of the residual energy ratio (solid) and its corresponding linear approximation (dashed) obtained by a least-mean square method.

In Figure 3.10 we provide the plot of the logarithm of the residual energy ratio $\log(r(n))$ for two different signals that vary in quality (a clean voice sample and a white noise sample) against the number of iterations of the MPA. We also plot

the best linear approximation $\log(\hat{r}(n))$ for both signals. As it is expected that the absolute value of the slope of the linear approximation is larger for the good quality signal ($RRF = 54 \cdot 10^{-2}$) compared to the poor quality signal ($RRF = 12 \cdot 10^{-2}$). Therefore, RRF can be considered as a correlate to speech quality. Although we have considered this quality feature in our work, it should be noted that it requires several runs of the MPA algorithm (at different iterations) before we can estimate the rate of decrease RRF.

3.3 Evaluation Method

In order to validate our proposed method for the non-intrusive voice quality estimation, we have tested our proposed objective quality features on the four databases described in Section 2.3. The MPA was implemented using the dictionary of Gabor atoms. We were motivated by studies which suggest that the Gabor dictionary of Gaussian functions can be a suitable atoms' basis to run the MPA on connected voice/speech samples [36, 45]. We have also run several tests using different other dictionaries such as Daubechies wavelets but the correlation results reported were moderate compared to the Gabor dictionary. The data were downsampled to 16 kHz and then analyzed. We evaluated the performance of our proposed algorithms using Pearson's correlation coefficient [62] which measures the linear dependence between the objective measures and the subjective voice quality ratings. Spearman rank correlation and sigmoidal mapping function were used too.

In order to understand the variations of our voice quality features MPA-PESQ and MPA-HASQI with respect to the number of iterations, we run our algorithms

on the four databases and computed the linear correlation coefficient as a function of the number of iterations for MPA-PESQ, MPA-HASQI and ECR features. As shown in Figure 3.11 , the maximum of correlations using the MPA-PESQ and MPA-HASQI features was obtained for a number of iterations ranging around to 18000-22000. Interestingly, the optimal numbers of iterations, which give the best correlation coefficient, for all databases (except D and D3 which shows moderate correlations) *almost* match. Note that, for the ECR feature, it seems that the best achieved correlation was reached at a lower number of iterations around 1000. This confirms the results reported by MacDonald *et al.* [45] for the ECR feature. Therefore, in the subsequent investigations, we have fixed the number of iterations at 20000 for the MPA-PESQ and MPA-HASQI algorithms while we chose the number of iterations to be equal to 1000 for the ECR feature.

3.4 Results

Metric	ρ	ρ_{spear}	ρ_{sig}	RMSE
VB	0.24	0.23	-0.33	0.36
HNR	-0.10	-0.10	-0.15	0.55
CPP	-0.26	-0.28	0.27	0.50
CPPs	-0.43	-0.41	-0.19	0.55
ITU-T P.563	-0.21	-0.27	-0.19	0.56
ECR	-0.42	-0.53	- 0.43	0.50
MPA-PESQ	-0.77	-0.77	-0.79	0.32
MPA-HASQI	-0.71	-0.70	-0.71	0.36
RRF	0.60	0.50	0.60	0.34

Table 3.1: Correlation values for different objective metrics for database D_1.

For comparison purpose, Tables 3.1, 3.2, 3.3, 3.4 show the obtained results, for

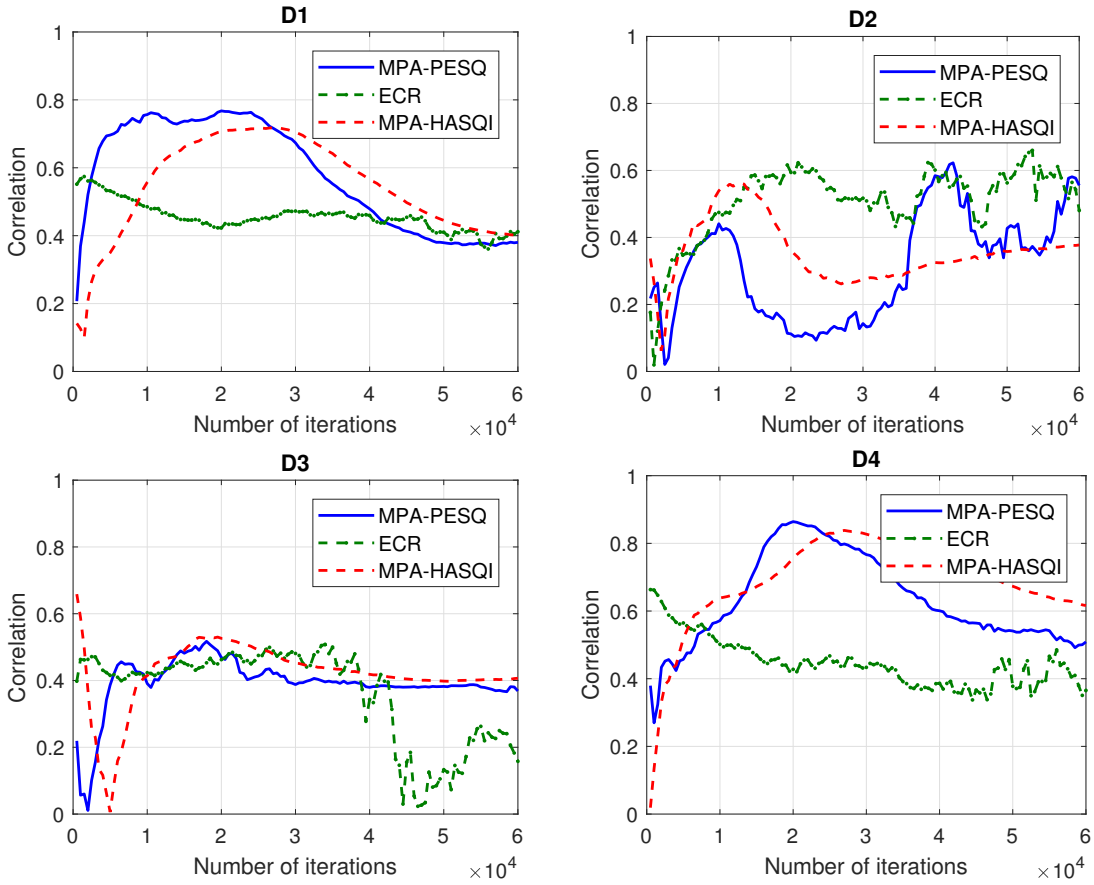

Figure 3.11: Correlation results for the voice samples of the all TE speaker databases D1, D2, D3, and D4

databases D1, D2, D3, and D4, when considering different objective speech quality algorithms. In particular, we have chosen to compare the performance of our algorithm (MPA-PESQ and MPA-HASQI) against: voice breaks (VB), harmonics-to-noise-ratio (HNR), cepstrum peak prominence (CPP), smoothed cepstrum peak prominence (CPPs), telephony standard ITU-T P.563, energy capture ratio (ECR) [50].

Figures 3.12, 3.13, 3.14 and 3.15 show the plots of the subjective ratings on the

Metric	ρ	ρ_{spear}	ρ_{sig}	RMSE
VB	0.16	0.19	-0.18	0.43
HNR	0.39	0.33	0.20	0.40
CPP	0.30	0.31	0.30	0.40
CPPs	0.38	0.42	0.38	0.30
ITU-T P.563	0.50	0.44	0.51	0.27
ECR	0.57	0.62	0.57	0.24
MPA-PESQ	0.21	0.20	0.23	0.50
MPA-HASQI	0.36	0.44	0.42	0.25
RRF	0.49	0.56	0.49	0.39

Table 3.2: Correlation values for different objective metrics for database D_2.

Metric	ρ	ρ_{spear}	ρ_{sig}	RMSE
VB	0.37	0.27	0.38	0.35
HNR	-0.51	-0.50	-0.20	0.51
CPP	-0.10	-0.14	-0.13	0.50
CPPs	-0.40	-0.22	-0.32	0.51
ITU-T P.563	-0.26	-0.22	-0.32	0.50
ECR	-0.47	-0.44	-0.47	0.35
MPA-PESQ	-0.47	-0.40	-0.40	0.43
MPA-HASQI	-0.51	-0.51	-0.40	0.39
RRF	-0.42	-0.34	-0.47	0.40

Table 3.3: Correlation values for different objective metrics for database $D3$.

Metric	ρ	ρ_{spear}	ρ_{sig}	RMSE
VB	0.23	0.27	-0.10	0.37
HNR	0.10	0.01	0.11	0.48
CPP	0.54	0.52	0.51	0.30
CPPs	0.02	0.06	0.10	0.40
ITU-T P.563	0.02	0.12	0.02	0.43
ECR	0.46	0.35	0.46	0.30
MPA-PESQ	0.86	0.75	0.87	0.20
MPA-HASQI	0.83	0.70	0.86	0.20
RRF	0.65	0.63	0.68	0.34

Table 3.4: Correlation values for different objective metrics for database D_4.

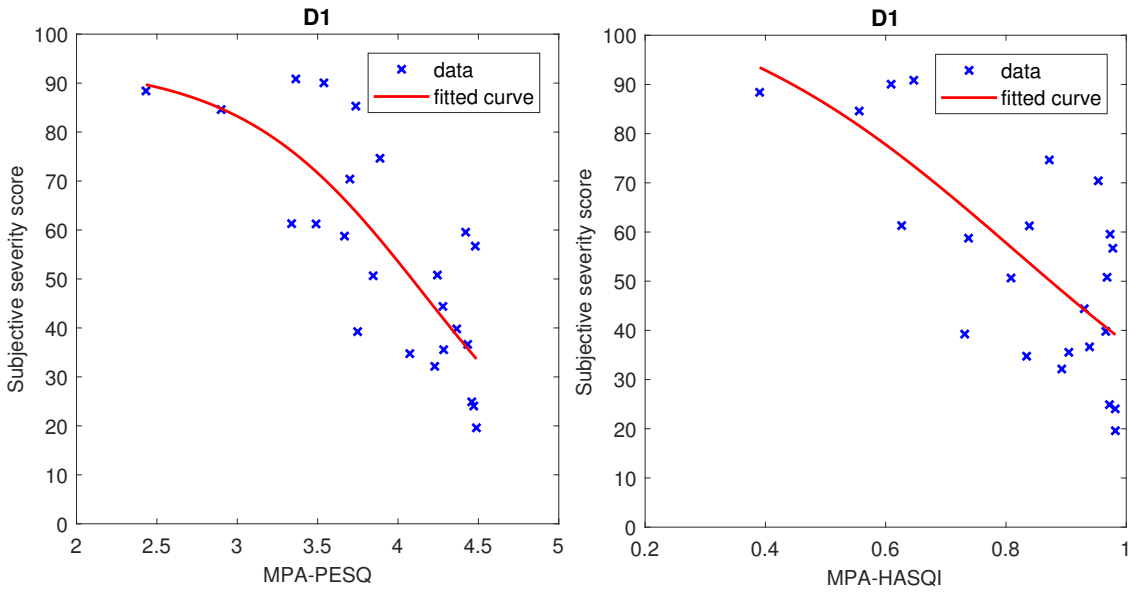

Figure 3.12: Subjective severity score for database D_1 and MPA-PESQ feature and MPA-HASQI feature.

X-axis against the objective ratings on the Y-axis for the two TE speech databases. These results show that our proposed MPA-PESQ and MPA-HASQI algorithm has a superior performance in measuring the quality compared to the other objective metrics. As it can be observed from the obtained results, the MPA-PESQ provides a correlation of 0.77 and 0.89 for D_1. For D_4, the correlation values are 0.75 and 0.87. Likewise, MPA-HASQI ranges from $0.70-0.71$ for D_1 and ranges from $0.70-0.86$ for D_4. The proposed algorithm has shown a superior performance as the MPA-PESQ went up to a range of 0.75 - 0.87, and $0.70-0.86$ for MPA-HASQI, respectively, when using a number of iterations that equals 20000.

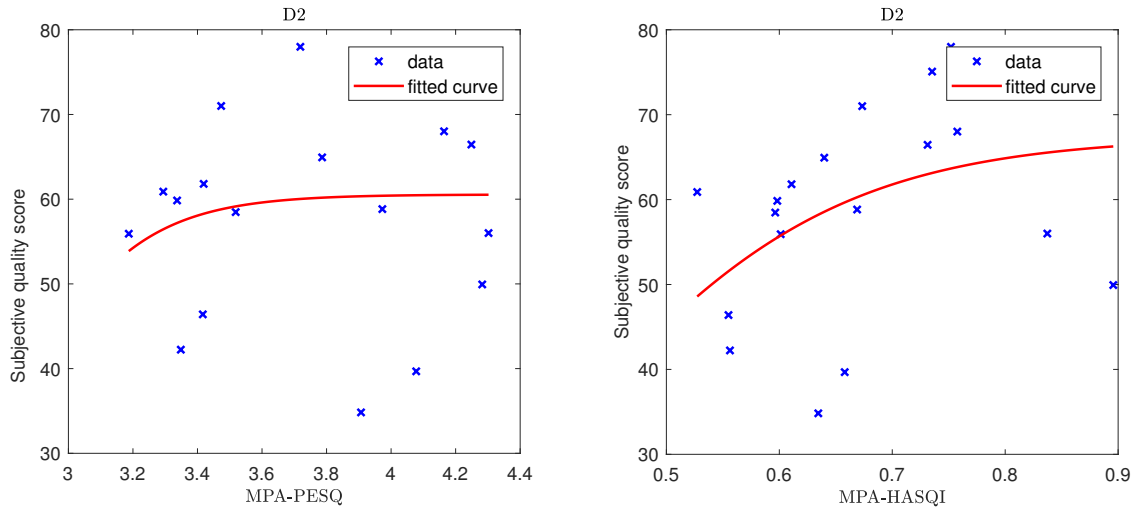

Figure 3.13: Subjective quality score for database $D2$ and MPA-PESQ, MPA-HASQI

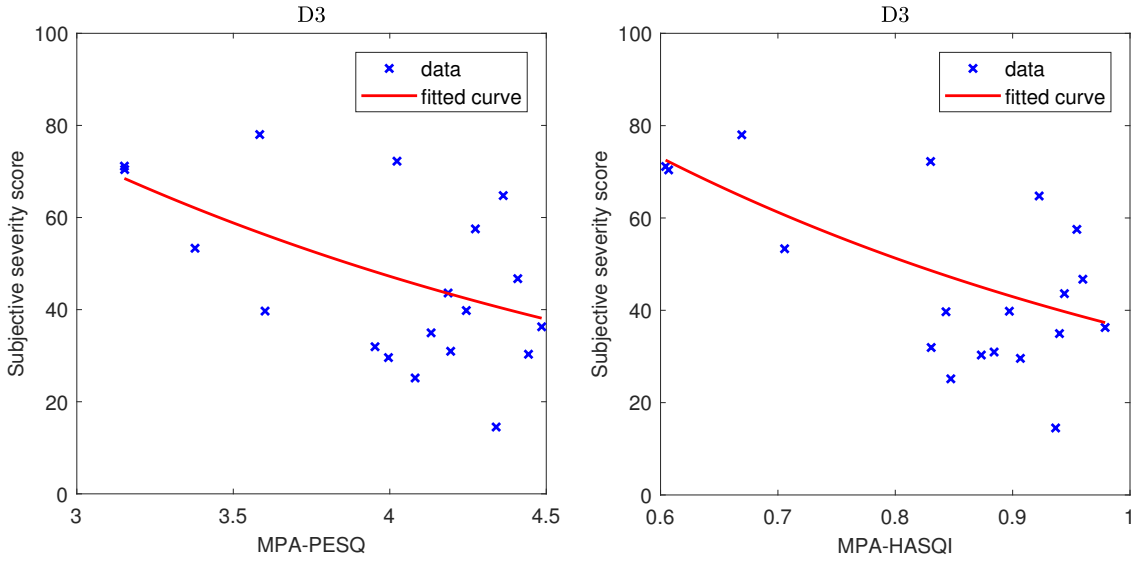

Figure 3.14: Subjective severity score for database $D3$ and MPA-PESQ, MPA-HASQI

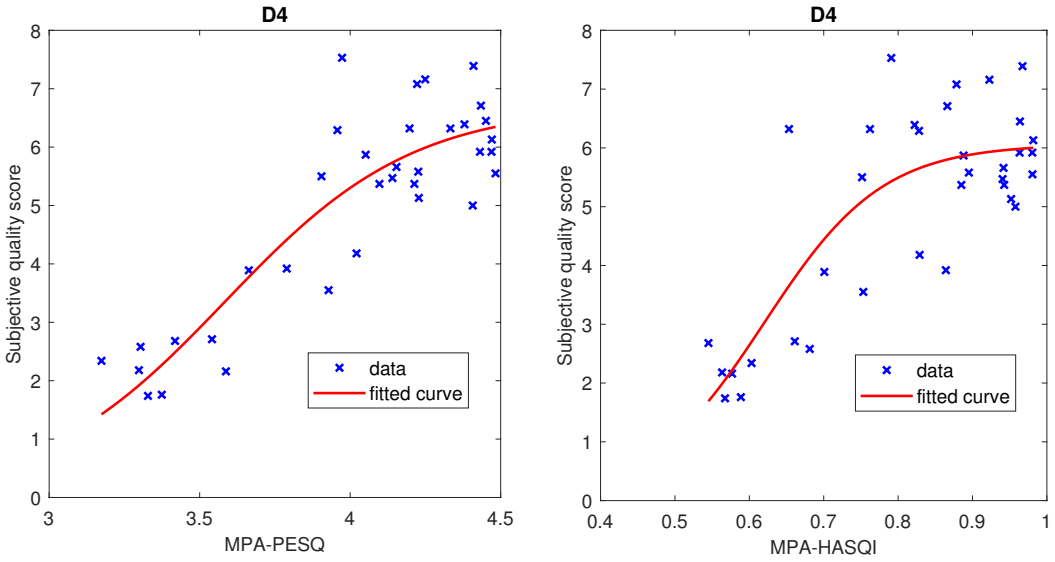

Figure 3.15: Subjective quality score for database $D4$ and MPA-PESQ, MPA-HASQI

3.5 Discussions

From a scientific viewpoint, it is necessary to provide some logical justification for the results obtained. More precisely, there is, up to now, no theoretical proof as to why the correlation obtained using MPA-PESQ or MPA-HASQI is maximal at certain number of iterations. We have plotted in the variations of the MPA-PESQ and MPA-HASQI scores against the number of iterations for all the samples from all databases, see Figures 3.16, 3.17. Interestingly, the results show that we have a maximum spread between the maximum obtained score and the minimum obtained score; further, this was reached at iterations very close to the number of iterations for the maximum correlation. We believe that this observation may lead to an explanation specific to the above paradigm. In fact, as we have a maximum spread between the scores of the voice samples assessed, there will be a better distinction between the quality of

the samples within a given databases and, therefore, a better correlation with the subjective quality ratings.

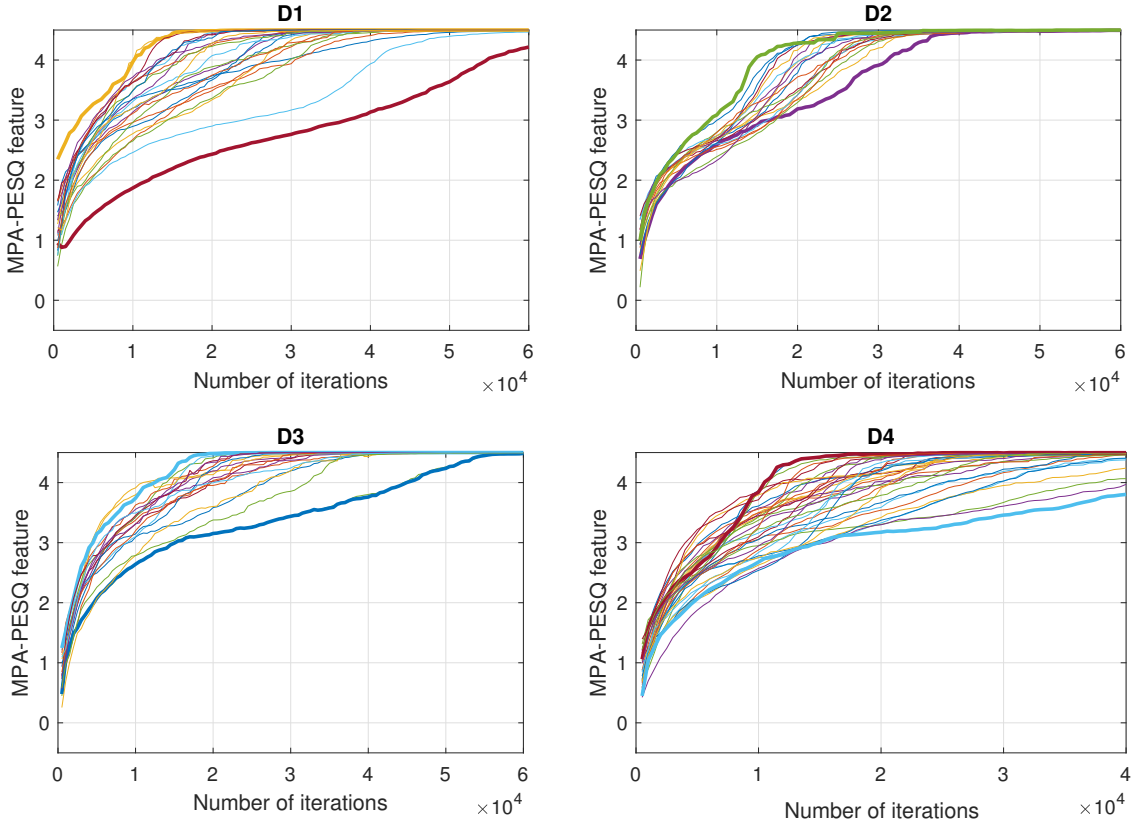

Figure 3.16: MPA-PESQ feature for the voice samples of the all TE speaker database with respect to the number of iterations of the MPA. The maximum score and minimum score samples are plotted in bold.

Moreover, it should be noted that the correlation results reported for D3 are moderate and this could be explained by many factors. First, it could be that the optimal choice of the dictionary of atoms for MPA for this database is not Gabor atoms, future extensive investigations maybe carried out to focus on the choice of the best atom given a dataset. Second, our logical assumption that distortions and

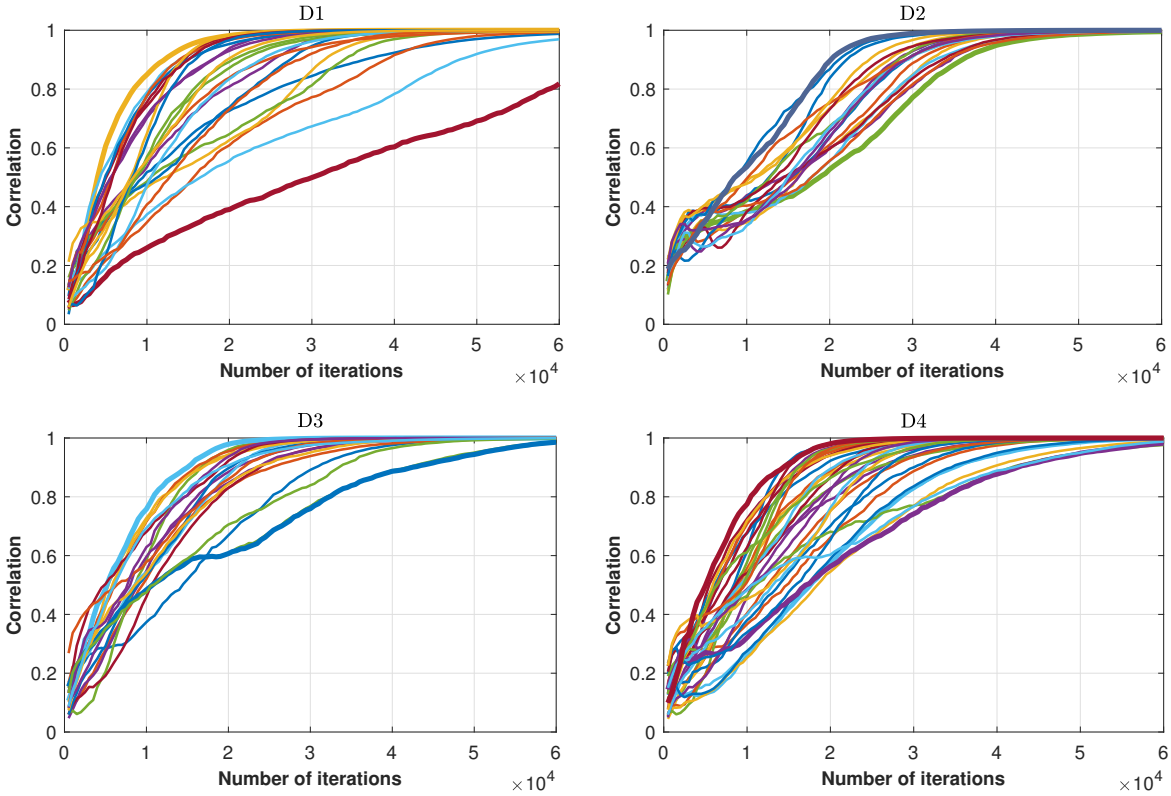

Figure 3.17: MPA-HASQI feature for the voice samples of the all TE speaker database with respect to the number of iterations of the MPA. The maximum score and minimum score samples are plotted in bold.

degradations in the TE speech sample are always modelled in the last iterations of the MPA algorithm might be not 100 percent true for some TE speakers. Finally, the instrusive algorithm PESQ and HASQI are not ideal and are often optimized and trained for some particular speech types. Therefore, the errors involved in the use of TE speech samples with the PESQ and HASQI might affect the final quality score.

3.6 Conclusion

In this chapter, we have proposed a new, non-intrusive perceptual estimation method for the quantification of disordered voice quality. Using the MPA and the Gabor dictionary of Gaussian time-frequency atoms, a reference sample signal was reconstructed from a given disordered voice sample. The traditional and widely used intrusive PESQ and HASQI standards were then used to predict the perceptual difference between the reference and degraded voice samples. The final score obtained is then used as an estimation of the subjective quality rating of disordered voice samples, in this case, those provide by TE speakers. The proposed technique was purposely tested on TE voice samples due to the fact that these samples are characterized by considerable levels of "noise" associated with the voice signal generated. The obtained scores were then correlated with quality ratings obtained from normal hearing listeners via auditory-perceptual evaluation methods. Our results showed a good correlation between the predicted and actual speech quality ratings using the proposed method. This correlation was significantly better than that achieved by conventional acoustic measures currently used in speech-language pathology at both research and clinical levels.

Chapter 4

Low Complexity Disordered Speech Quality Estimation

4.1 Introduction

Voice and speech quality estimation is an important topic of research with many applications in telecommunication and biomedical engineering. We called it Low Complexity Disordered Speech Quality Estimation LCDSQE, because the acoustic features extracted are simply based on LP analysis and do not require advanced tools such as PESQ and HASQI. Also the regression model is linear and simple to implement compared to deep learning model for example. In this chapter, our goal is to propose acoustical features which are easily extracted (computationally simple) from a given speech signal and which are shown to correlate well with subjective scores of TE speech. First, the voiced frames of the acoustical speech signals are extracted using the autocorrelation method [63] and the corresponding pitch estimation per

voiced frame is obtained. The voiced frames of the speech are evaluated using an 18-th order LP analysis based on the Levinson-Durbin algorithm. Speech quality features are extracted by computing the average over all frames of high order statistics (mean, standard deviation, skewness and kurtosis) of the LP coefficients, the cepstral coefficients and the LP residual signal. Furthermore, a vocal tract model has been extracted for each voiced frame by computing the parameters of an acoustical tube formed by interconnecting 18 uniform cross sectional tubes. The vocal tract parameters yielded extra speech quality features. Finally, the extracted speech quality features have been used to train and test different support vector machine models on the dataset D4 (described in Section 2.3) that contains 35 TE speech samples.

4.2 Speech quality evaluation method

Here our proposed approach for extracting speech quality features from disordered speech signals consists of three main stages. First, preprocessing is conducted to detect voiced and unvoiced speech frames. We use a temporal approach based on the autocorrelation method. Then, LP analysis is performed to extract the LP coefficients, the cepstral coefficients and the residual signal from each frame marked as voiced by the first preprocessing stage.

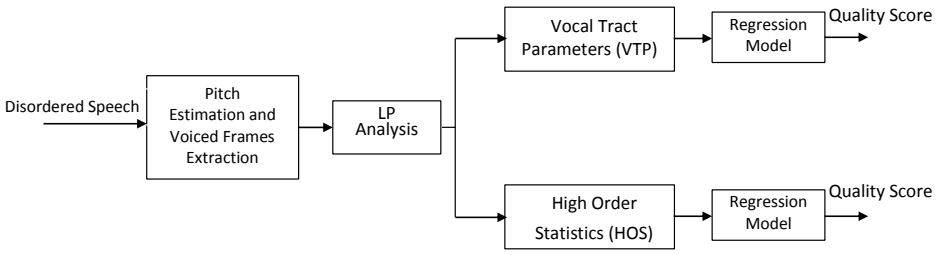

Figure 4.1: The proposed speech quality algorithm.

The LP coefficients are used to derive a vocal tract model by calculating the reflection and the cross sectional areas of the acoustic tube model which provides the first group of acoustic features. Besides, higher order statics are obtained from LP analysis coefficients and residual signal which constitute the second group of acoustical features. Each group of features is used in a regression-based mapping to provide quality scores for the disordered speech signals. The schematic of the proposed method for speech quality estimation is depicted in Figure 4.1. The different stages listed above are detailed in the next subsections.

4.2.1 Pitch Period Estimation and Voiced Frames Extraction

Pathological speech signals are different in terms of their pitch period estimates. It is suggested that inclusion of pitch average estimates in computational models for voice quality may help improve the accuracy of these models. In non-intrusive speech quality measurement algorithms, such as the ITU standard P.563 and the Low-Complexity Nonintrusive Speech Quality Assessment (LCQA) proposed in [64], pitch is used as a feature for quality assessment. We use the autocorrelation method to provide an estimate of the pitch length for the frames marked as voiced. The speech signal is divided into 20-ms frames with 50% overlap using the Hann window. The autocorrelation function is then calculated and normalized for each 20-ms frame. The current n-th speech frame is marked as voiced when the second maximum peak of the normalized autocorrelation exceeds 0.5. This extraction method is summarized in Fig. 4.2. The corresponding pitch length $T(n)$ is obtained by computing the time distance from the origin to the peak.

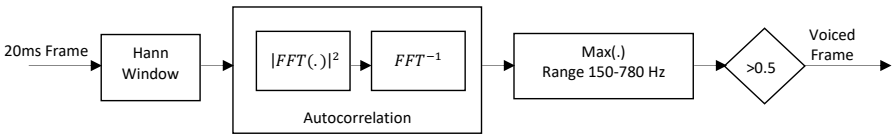

Figure 4.2: Pitch period estimation and voiced frames extraction method using the autocorrelation method.

4.2.2 Linear Prediction Analysis

As the degree of severity of dysphonia becomes higher, the speech signal tends to have more and more aperiodic, irregular and noncoherent components. This has been observed in [65] for pathological voices in sustained vowels. The linear prediction (LP) analysis performed in [65] has been used to derive high order statistics (skewness and kurtosis) from the LP residual signal from each frame of the sustained vowel signal. Since continuous pathological voices may contain voiced and/or unvoiced frames, we propose to perform the LP analysis only on voiced frames. In fact, voiced frames are quite quasi-periodic signals which suggests to use an Auto Regressive (AR) filter to model the production of each speech frame.

The Levinson-Durbin algorithm is used to derive an 18th-order all pole LP model for each 20-ms frame marked as voiced by the preprocessing done in Section (4.2.1). The model is characterized by a set of 18 LP coefficients $\{a_i(n)\}_{1\leq i\leq 18}$ where n denotes the frame number.

4.2.2.1 Cepstral Coefficients

Cepstral coefficients are the coefficients of the inverse Fourier transform representation of the log magnitude of the spectrum of the signal. Once LP coefficients are obtained, it is possible to directly extract cepstral coefficients from them. Assume we want to

extract $p < 18$ cepstral coefficients from the obtained 18 LP coefficients $\{a_i(n)\}_{1 \leq i \leq 18}$ then we use the following formula:

$$c_i(n) = a_i(n) + \sum_{l=1}^{i-1} \frac{l}{i} c_l(n) a_{i-l}(n), \quad 2 \leq i < p \tag{4.1}$$

where $c_1(n) = r_{xx}(0)$ representing the maximum autocorrelation of the n-th frame of the speech signal. In this work we extracted $p = 5$ cepstral coefficients per frame.

4.2.2.2 LP Residual

LP residual may bring information on the abnormal behaviour of the speech production system (vocal folds, vocal tract, turbulence noise...etc) which could be used for disordered speech quality assessment [65]. LP residual represents the error between the original signal and the synthesized (estimated) signal using the derived LP coefficients. The residual of the LP analysis for the n-th voiced frame is obtained as

$$e_n(k) = x_n(k) - \sum_{i=1}^{18} a_n(i) x_n(k-i) \tag{4.2}$$

where $x_n(k)$ represents the value of the original signal at the k-th sample of the n-th frame. Once the LP analysis has been performed on each voiced frame of the speech signal, we derive different quality features as detailed in Section 4.2.4

4.2.3 Vocal Tract Modelling

This speech assessment block focuses on the speech production system. The human voice production system is composed of an air pressure source (lungs), a modulator

(vocal folds) and a resonating system (vocal tract). Airflow created by the lungs excites the vocal chords to generate either a voiced sound or an unvoiced sound (also called voiceless sound). During voiced sounds, a low-frequency (quasi-periodic) sound is generated. The vocal tract acts as a filter that shapes the spectral content of the sound. Controlled contractions and relaxations of the vocal tract muscles change the shape of the vocal tract, and thus its resonant frequencies, to produce the different voiced sounds. During unvoiced sounds, a turbulent, aperiodic excitation is created by forcing air through a constriction in the vocal tract, for example when the upper teeth are placed on the lower lip.

In [58], vocal tract models are used to design a non-intrusive speech quality assessment method that was later implemented in the ITU-T P.563 standard used in telecommunications [44]. The idea is to model the vocal tract as a set of acoustic tubes (with uniform cross-section area) arranged in a series configuration, see Figure 4.3. Each tube has a different section area that changes over time. The idea is to use LP model coefficients to extract the reflection coefficients and the tube section areas for voiced speech frames. The number of tubes is equal to the order of the LP (number of LP coefficients). In [44], the vocal tract is modelled as eight concatenated acoustic tubes which is suitable for narrowband signals sampled at around 8 kHz. In our work, we model the vocal tract using a series of 18 acoustic tubes (LP order equals 18) which is suitable for wideband signals associated with disordered speeche. Note that although TE speech is produced by patients with *no vocal tract* since the larynx is completely removed, here our intention is to model *how good* the TE speech is produced. Therefore, since *good* speech is naturally produced by normal people (with vocal tract), this justifies our approach in using a vocal tract model to extract

TE speech quality features.

For each voiced frame of the signal, the reflection coefficients are calculated from the LP coefficients using the following recursion:

$$r_i(n) = \alpha_{i,i}(n), \qquad 1 \leq i \leq 18 \tag{4.3}$$

$$\alpha_{i-1,l}(n) = \frac{\alpha_{i,l}(n) - r_i(n)\alpha_{i,i-l}(n)}{1 - r_i(n)^2}, \qquad 1 \leq l < i \tag{4.4}$$

such that $\alpha_{18,i} = a_i(n)$ corresponding to the i-th coefficient for the LP model of the n-th frame. Once the reflection coefficients $\{r_i(n)\}_{1 \leq i \leq 18}$ are extracted, the cross section areas can be computed using the recursion [44]:

$$S_i(n) = \frac{1 + r_i(n)}{1 - r_i(n)} S_{i+1}(n), \quad i = 18, 17, \cdots, 1. \tag{4.5}$$

The cross section area S_{18} can be obtained by letting $S_{19} = 1$.

4.2.4 Features Extracted

Based on the above LP analysis and vocal tract modelling, we derive two groups of features which will allow us to assess the speech quality of our disordered speech samples.

4.2.4.1 Higher Order Statistics

High order statistics (HOS) analysis has been used in classification of pathological voices [66] and in robust voice activity detection [67] with very promising results. It has the advantage of not requiring a periodic or quasiperiodic voice signal to permit

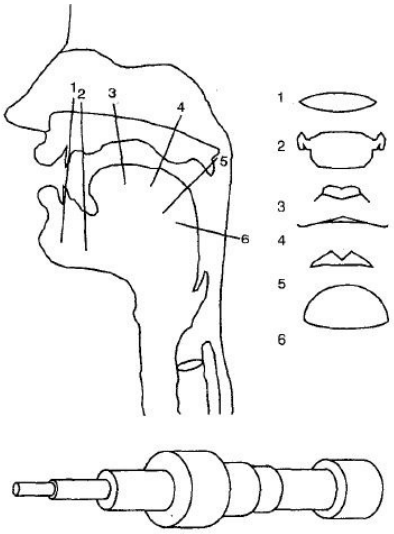

Figure 4.3: Illustration of the vocal tract uniform-cross-sectional-area tube model [58] Top: true cross-section shapes of the vocal tract sketched at different locations. Bottom: a simplified uniform-cross-sectional-area tube model (with 8 tubes) of the vocal tract. In this work we consider a tube model with 18 acoustic tubes

.

a reliable analysis.

In this work, we derive 12 HOS for each frame of the speech signal by considering the 4 HOS (mean, variance, skewness and kurtosis) of the LP coefficients $\{a_i(n)\}_{1\leq i\leq 18}$, the cepstral coefficients $\{c_i(n)\}_{1\leq i\leq 5}$ and the LP residual signal $\{e_i(n)\}_{1\leq i\leq N}$ where N is the number of speech samples within one frame and n is the corresponding frame index. The 12 HOS statistics are averaged across all the voiced frames to yield the features $\texttt{HOS}_1$,...,$\texttt{HOS}_{12}$ which are defined according to Table 4.1.

To this group of features, we add the $\texttt{HOS}_{13}$ feature which is computed by taking the average of the different pitch lengths $T(n)$ for all the voiced speech frames. Also the number of voiced frames is taken as a quality feature and denoted $\texttt{HOS}_{14}$.

To illustrate the dependence of these high order statics on the speech quality, we consider the mean value of the LP coefficients, denoted $\mu_a(n)$, for the n-th frame.

	Mean μ_x	Variance σ_x	Skewness γ_x	Kurtosis κ_x
LP coefficients	$\texttt{HOS}_1$	$\texttt{HOS}_2$	$\texttt{HOS}_3$	$\texttt{HOS}_4$
CP coefficients	$\texttt{HOS}_5$	$\texttt{HOS}_6$	$\texttt{HOS}_7$	$\texttt{HOS}_8$
LP residual	$\texttt{HOS}_9$	$\texttt{HOS}_{10}$	$\texttt{HOS}_{11}$	$\texttt{HOS}_{12}$

Table 4.1: High-order statistics (HOS) features.

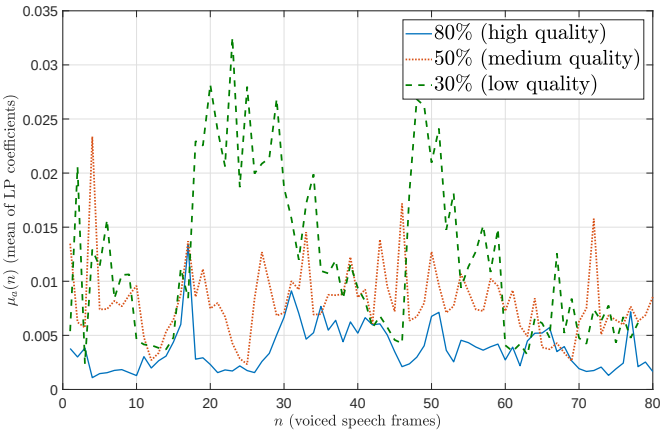

Figure 4.4: Average value of LP coefficients for each voiced TE speech frame.

The transfer function of the all poles LP model, for a given frame is given by

$$H_n(z) = \frac{1}{1 + \sum_{i=1}^{18} a_n(i) z^{-i}}. \tag{4.6}$$

Therefore, one has

$$\mu_a(n) = \frac{1}{18} \sum_{i=1}^{18} a_n(i) = \frac{1 - H_n(1)}{18 H_n(1)}. \tag{4.7}$$

This implies that the mean of the LP coefficients $\mu_a(n)$ will increase as the value of the DC-gain $H_n(1)$ decreases. For disordered TE speech samples, it is observed that

the voiced segments of the speech produced by TE patients will tend to have a gain attenuation (lower values of $H_n(1)$) as the quality of the speech signal gets worse (see Figure 4.4). Therefore, the average of $\mu_a(n)$ across all frames is likely to be inversely proportional to the overall quality of the speech.

4.2.4.2 Vocal Tract Parameters

The second group of speech quality features is based on the vocal tract modelling done in Section 4.2.3. To extract speech quality features from the instantaneous vocal tract tubes model we use the idea that, due to the removal of the larynx, TE speech can be thought to have an "imperfect" speech production system that mimics an abnormal vocal tract configuration. In this work we wanted to extract as many voice features as possible. We consider the maximum, minimum and average of each cross section area along the whole speech which results in $18 \times 3 = 54$ different features. These features were assigned the labels $\texttt{VTP}_1$,...,$\texttt{VTP}_{54}$ and are defined as follows:

$$\texttt{VTP}_i = \max_n(S_i(n)) \tag{4.8}$$

$$\texttt{VTP}_{i+18} = \min_n(S_i(n)) \tag{4.9}$$

$$\texttt{VTP}_{i+36} = \text{avg}_n(S_i(n)) \tag{4.10}$$

for $i \in \{1, \cdots, 18\}$. The extracted features are then fed to different models which are fitted and compared using advanced regression analysis performed on a TE disordered speech database as detailed in the next section.

4.3 Speech database

To train the different regression models, we use dataset D4 (described in Section 2.3) which is a database of 35 tracheoesophageal (TE) speech recordings. The motivation behind choosing D4 over the other datasets we have, is the fact that this dataset contains the largest number of TE speech samples. This is important to train a fairly good regression model and to be able to divide the dataset into training and test datasets. The other TE speech datasets are left to test the robustness of the obtained regression model against other TE speech samples collected at different times. Note that optimally, we would need to collect a larger dataset to train an accurate model; however this exceeds the scope of the thesis.

4.4 Results and discussions

The features extracted from the vocal tract modelling ($\texttt{VTP}_1, \cdots, \texttt{VTP}_{54}$) and from the higher-order statistics ($\texttt{HOS}_1, \cdots, \texttt{HOS}_{14}$) are used to train different regression models. First, for each group of features, forward stepwise regression (FSR) [68] is performed to prioritize the features within the group. Initially no predictors are included in the model. Then, as a first step, we check all the possible models with one predictor against the coefficient of determination R^2 (R squared)

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{4.11}$$

where the y_i's are the subjective scores (true observations), $\hat{y}_i$'s are the estimation scores and $\bar{y}$ is the mean value of the y_i's data. Then, the feature that gives a model

with the highest R^2 is retained. The second step consists in checking all the models with two features by adding another feature to the previously selected feature. This procedure is repeated until we select all the available features. Note that the FSR algorithm stops also if, otherwise, the value of R^2 reached 1 where in this case the remaining features are discarded. Finally, we obtain a natural ordering of the features by their importance. These results are provided in Table 4.2.

For example, if we want to use a model with 3 HOS features then the best set of 3 features (from the set of 14 features) is $\texttt{HOS}_5, \texttt{HOS}_9, \texttt{HOS}_{11}$. Similarly, a model with 3 VTP features would contain $\texttt{VTP}_5, \texttt{VTP}_{20}$ and $\texttt{VTP}_4$. Note that the FSR algorithm has stopped after selecting 34 features (out of 54 features) because the value of R^2 reached 1 and the addition of any other features will not bring further information.

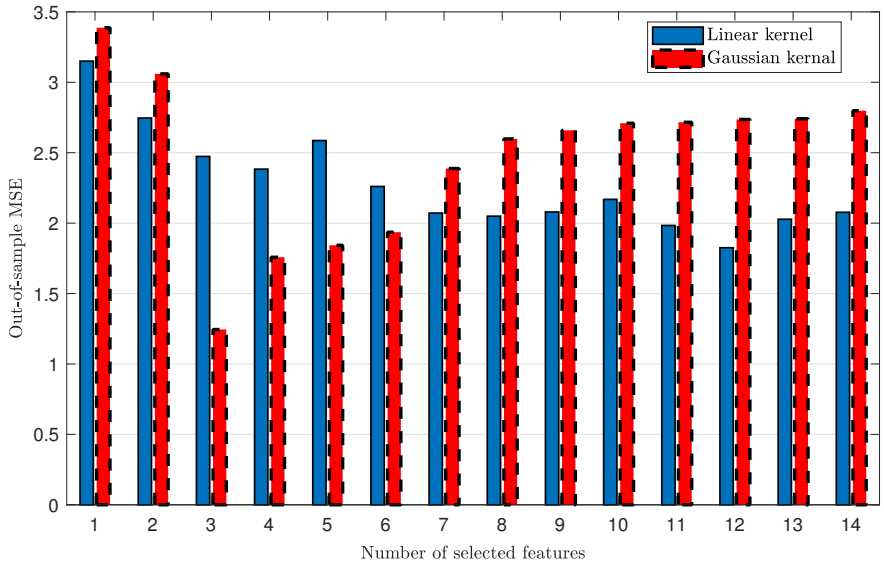

Figure 4.5: Feature selection from the HOS statistics group

Then, we use K-folds cross validation method [69] to select the best set of features that guarantees the lowest prediction error (test error). This allows to avoid the prob-

R^2	Added HOS feature
0.1469	$\texttt{HOS}_{11}$
0.260	$\texttt{HOS}_{9}$
0.435	$\texttt{HOS}_{5}$
0.473	$\texttt{HOS}_{2}$
0.546	$\texttt{HOS}_{13}$
0.560	$\texttt{HOS}_{10}$
0.657	$\texttt{HOS}_{1}$
0.679	$\texttt{HOS}_{3}$
0.708	$\texttt{HOS}_{6}$
0.731	$\texttt{HOS}_{7}$
0.744	$\texttt{HOS}_{4}$
0.761	$\texttt{HOS}_{12}$
0.772	$\texttt{HOS}_{14}$
0.803	$\texttt{HOS}_{8}$

R^2	Added VTP feature
0.382	$\texttt{VTP}_{5}$
0.480	$\texttt{VTP}_{20}$
0.560	$\texttt{VTP}_{4}$
0.663	$\texttt{VTP}_{38}$
0.712	$\texttt{VTP}_{18}$
0.739	$\texttt{VTP}_{24}$
0.762	$\texttt{VTP}_{8}$
0.783	$\texttt{VTP}_{22}$
0.804	$\texttt{VTP}_{21}$
0.831	$\texttt{VTP}_{35}$
0.860	$\texttt{VTP}_{34}$
0.870	$\texttt{VTP}_{54}$
0.878	$\texttt{VTP}_{28}$
0.886	$\texttt{VTP}_{36}$
0.893	$\texttt{VTP}_{50}$
0.900	$\texttt{VTP}_{51}$
0.926	$\texttt{VTP}_{46}$
0.939	$\texttt{VTP}_{14}$
0.947	$\texttt{VTP}_{17}$
0.965	$\texttt{VTP}_{33}$
0.973	$\texttt{VTP}_{7}$
0.980	$\texttt{VTP}_{6}$
0.983	$\texttt{VTP}_{26}$
0.988	$\texttt{VTP}_{37}$
0.990	$\texttt{VTP}_{53}$
0.991	$\texttt{VTP}_{1}$
0.993	$\texttt{VTP}_{39}$
0.994	$\texttt{VTP}_{23}$
0.998	$\texttt{VTP}_{9}$
0.999	$\texttt{VTP}_{3}$
0.999	$\texttt{VTP}_{49}$
0.999	$\texttt{VTP}_{52}$
0.999	$\texttt{VTP}_{47}$
1	$\texttt{VTP}_{2}$

Table 4.2: Forward stepwise regression results.

	Model	Selected Features
HOS Statistics	Linear	$\texttt{HOS}_1$, $\texttt{HOS}_2$, $\texttt{HOS}_3$, $\texttt{HOS}_4$, $\texttt{HOS}_5$, $\texttt{HOS}_6$, $\texttt{HOS}_7$, $\texttt{HOS}_9$, $\texttt{HOS}_{10}$, $\texttt{HOS}_{11}$, $\texttt{HOS}_{12}$, $\texttt{HOS}_{13}$
	Gaussian	$\texttt{HOS}_5$, $\texttt{HOS}_9$, $\texttt{HOS}_{11}$
VTP Parameters	Linear	$\texttt{VTP}_4$, $\texttt{VTP}_5$, $\texttt{VTP}_8$, $\texttt{VTP}_{18}$, $\texttt{VTP}_{20}$, $\texttt{VTP}_{21}$, $\texttt{VTP}_{22}$, $\texttt{VTP}_{24}$ $\texttt{VTP}_{34}$, $\texttt{VTP}_{35}$, $\texttt{VTP}_{38}$, $\texttt{VTP}_{54}$
	Gaussian	$\texttt{VTP}_4$, $\texttt{VTP}_5$, $\texttt{VTP}_{20}$, $\texttt{VTP}_{38}$

Table 4.3: Selected features for each model.

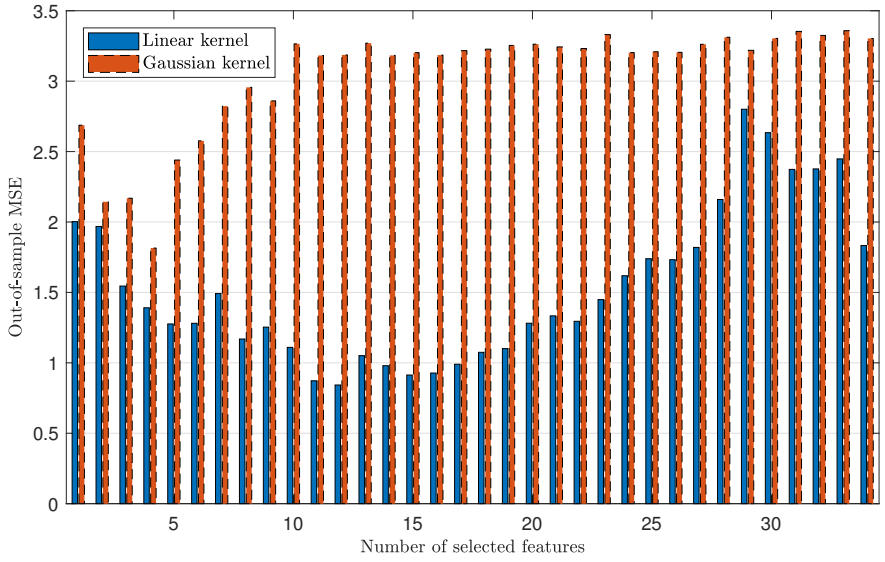

Figure 4.6: Feature selection from VTP Parameters group

lem of overfitting. For each number of selected features (obtained from the FSR), we use a 7-fold cross validation by training and testing support vector machines regression models [70] with two different kernel functions: linear and Gaussian. Fig. 4.5 and Fig.4.6 plot the out-of-sample mean square error (MSE) for each cross-validated model resulted from the selected features for the HOS predictors group and the VTP predictors group, respectively. From these figures we can determine the set of features from each group that minimizes the out-of-sample MSE. These sets of features are

given in Table 4.3 for each group and each kernel function.

Table 4.4: Correlation values of the proposed objective metrics.

Metric	Correlation (training dataset)	Correlation (test dataset)
HOS-Linear	0.89	0.78
HOS-Gaussian	0.78	0.63
VTP-Linear	0.93	0.84
VTP-Gaussian	0.98	0.70

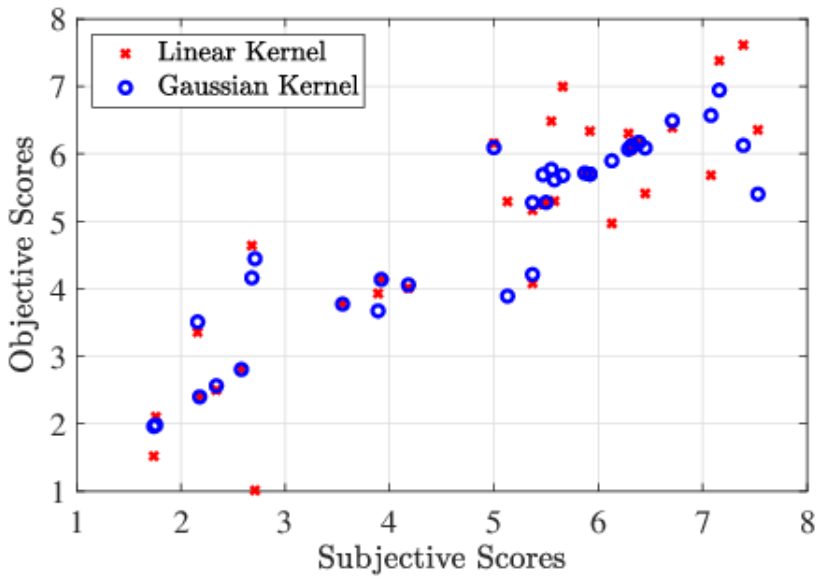

Figure 4.7: Scatter plot of subjective scores against the objective scores derived from the VTP parameters-based model.

Once, the sets of features are selected, each set of features is used to train a model (linear or Gaussian). The data set consists of 35 recordings and is divided into two separate groups. The first group contains 25 recordings and serves as a training set to train the regression model, while the other 10 recordings are used to test the prediction

capabilities of this regression model. The performance of our proposed algorithms is evaluated using Pearson's correlation coefficient. Table 4.4 shows the results obtained

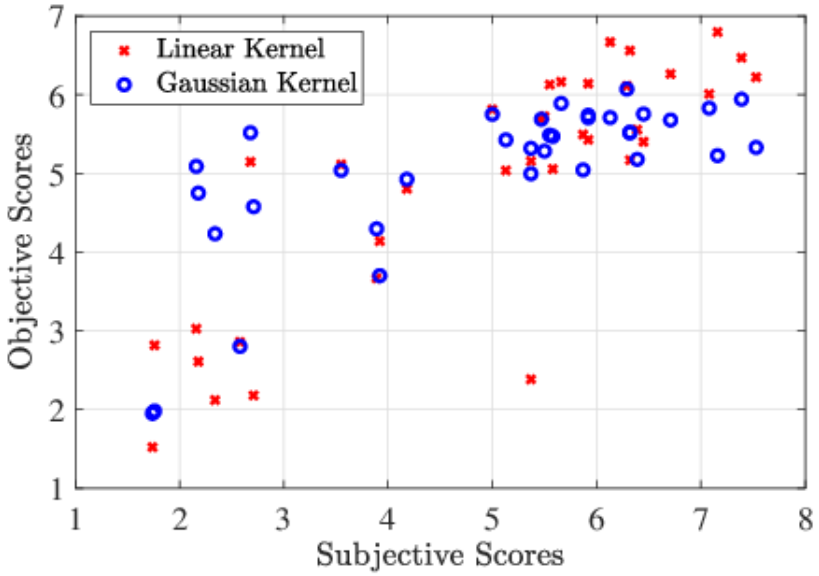

Figure 4.8: Scatter plot of subjective scores against the objective scores derived from the HOS statistics-based model.

from the proposed objective metrics. Applying support vector regression (SVR) with linear kernel to the selected HOS features yields a correlation of 0.89 with the training dataset samples, while gives 0.78 correlation with the test dataset. Using the SVR technique with a Gaussian kernel to get an objective model for the selected HOS features has a slightly weaker performance in terms of prediction capabilities and overfitting avoidance. The correlation values are 0.78 and 0.63 for the training and the test datasts respectively. Applying SVR model with a linear kernel to the vocal tract VTP features led to a better performance in terms of overfitting avoidance and bias minimization. The correlation values for the traininig and the test datasets were 0.93 and 0.84. Changing the kernel to Gaussian has increases the training correlation

to 0.98 while decreasing the testing correlation to 0.70. Fig. 4.8 shows the scatter plot of objective scores against subjective scores for the each of VTP- and HOS-based metrics. These results suggest that an SVR model with linear kernel would perform better than an SVR model with Gaussian kernel although the latter uses less number of features as shown in Table 4.3. Also, the VTP-based models have performed slightly better than the HOS-based features which shows that features extracted from the vocal tract modelling (speech production system) consist of good predictors for disordered speech.

The obtained correlation results for the proposed algorithms are much better than the correlation obtained from previously proposed features in the litterature such as the Harmonicto-Noise-Ratio (HNR), Cepstral Peak Prominence (CPP), the ITU-T recommendation P.563 amongst others, see Table 4.5. However, the obtained correlation values when using these regression models for the other TE speech databases were moderate. This can be explained by the fact that the number of samples (in D4) used to train the models is not enough to capture all the different TE speech distortions and inconsistencies. Also, the fact that the datasets were collected at different conditions, rated by different listeners for different speech quality measures, greatly affects the discrepancies between the acoustical features of each dataset.

4.5 Conclusion

This chapter introduces a new nonintrusive algorithm, with low computational complexity, suitable for disordered speech quality estimation. Using an 18-order LP analysis applied to voiced frames of the acoustic speech signal, we derived up to

	Algorithm	Correlation
D1	HOS-Linear	-0.48
	HOS-Gaussian	-0.46
	VTP-Linear	-0.49
	VTP-Gaussian	-0.56
D2	HOS-Linear	0.41
	HOS-Gaussian	0.47
	VTP-Linear	0.53
	VTP-Gaussian	0.60
D3	HOS-Linear	-0.64
	HOS-Gaussian	-0.37
	VTP-Linear	-0.56
	VTP-Gaussian	-0.50
D4	HOS-Linear	0.85
	HOS-Gaussian	0.73
	VTP-Linear	0.90
	VTP-Gaussian	0.91

Table 4.5: Comparison of the correlation values obtained using different quality estimation methods.

14 high-order statistical (HOS) based features and 54 vocal tract parameters (VTP) based features. We used a set of 35 TE speech samples to train different support vector regression models after performing features selection using forward stepwise regression and K-folds cross validation. The obtained models are shown to be able to predict the quality scores of the subjective scores with a correlation coefficient than ranges from 0.78 to 0.98 for the training dataset and from 0.63 to 0.84 for the test dataset. The obtained results of this work suggest that the HOS and VTP features, which are extracted from a simple LP analysis of the acoustic speech signal, can be a cheap alternative to the more complex existing non intrusive algorithms for quality estimation of pathological speech. However, given the small number of TE speech samples used to train the regression models, the proposed low complexity approach

did not correlate well when tested with other TE speech datasets.

Chapter 5

Deep Learning-Based Quality Assessment for Tracheoesophageal Speech

5.1 Introduction

In this chapter we aim to develop a non-intrusive speech quality estimation algorithm optimized for TE speech samples which is computationally simple and robust across databases. Deep neural networks (DNN) drew a massive amount of attention in the recent years. DNNs are employed in many applications such as autonomous vehicles, image processing, natural language processing (NLP), and automatic speech recognition (ASR) [71]. In this chapter we propose a speech quality estimation algorithm that exploits the advances in deep learning (DL) to apply it to a group of suitably extracted features that are shown to correlate well with disordered speech quality.

More specifically, we extract 154 acoustic speech quality features which are cut down to 60 features and then used as inputs to a deep neural network with three layers. The neural network is trained using a newly generated dataset by artificially adding different noise levels to an existing normal quality speech dataset. Then, the HASQI algorithm of [48] is used to provide objective scores for the generated dataset which is used for the training of the neural network. Once the neural network is trained, it is validated on our four TE speech datasets described in Section 2.3 and shows a high correlation values with subjective scores.

5.2 Speech Quality Evaluation Method

Objective voice and speech quality evaluation methods are classified into two main categories: intrusive and non-intrusive. For intrusive methods, the evaluation algorithm needs access to a high quality reference signal, in order to infer the quality of the signal under test. However, in disordered speaker populations such as those who produce TE speech, a clean reference signal is unavailable. Our proposed speech quality estimation method in this chapter is based on the extraction of different features from the speech recordings and then mapping the extracted features to a predicted quality score using a DNN. The proposed features are classified into four groups: linear prediction-based higher order statistics (HOS), vocal tract parameters (VTPs), Mel Frequency Cepstrum Coefficients (MFCCs), and Gammatone Frequency Cepstrum Coefficients (GFCCs). The extraction algorithms of these features are detailed in the following subsections.

5.2.1 Linear Prediction-Based Higher Order Statistics

Following our work in Chapter 4, we first used the autocorrelation method to provide an estimate of the pitch period for the frames marked as voiced. The speech signal was divided into 20-ms frames with 50% overlap using the Hann window. The autocorrelation function was then calculated and normalized for each 20-ms frame. The current speech frame was marked as voiced when the second maximum peak of the normalized autocorrelation exceeded 0.5. The corresponding pitch period length was then obtained by computing the temporal distance from the origin to the peak. The Levinson-Durbin algorithm [72] was then used to derive an 18th-order all pole LP model for each 20-ms voiced frame. Once the LP coefficients were obtained, we directly extracted five cepstral coefficients from them and the corresponding LP residual signal. Finally, for each voiced frame, we extracted 12 high order statistics (HOS) corresponding to 4 HOS (mean, variance, skewness and kurtosis) for each LP coefficients, cepstral coefficients and LP residual. The 12 HOS statistics were averaged across all the voiced frames to yield the final feature set. To this group, we added two additional features: 1) the average of the pitch lengths and 2) the number of voiced frames yielding a total of 14 features.

5.2.2 Vocal Tract Parameters

The second group of voice and speech quality features was based on a vocal tract modelling similar to [44]. The idea is to model the vocal tract as a set of acoustic tubes (with a uniform cross-section area, see Fig. 4.3) arranged in a serial configuration, and to extract the reflection coefficients and the tube sectional areas for voiced speech frames using the LP coefficients. We modelled the vocal tract using a series of 18

acoustic tubes (LP order equals 18) which is suitable for wideband signals associated to disordered samples. For each voiced frame of the signal, the reflection coefficients were calculated from the LP coefficients which allowed for computation of the cross-sectional areas. Then, we considered the maximum, minimum and average of each cross section along the whole speech which resulted in $18 \times 3 = 54$ different features.

5.2.3 Mel Frequency Cepstrum Coefficients (MFCCs)

MFCC is defined as the real cepstrum of a windowed short-time signal derived from the FFT of that speech signal [73]. This group of features was extracted from transforming the short-term disordered speech spectra into the nonlinear mel scale [74] using the formula

$$m = 2595 \log_{10}(1 + f/100). \tag{5.1}$$

The mel scale, is a perceptual scale of pitch values judged by listeners to be equal in distance from one another. The input speech signal was framed into 256 samples each, with each frame partitioned using a Hamming window and transformed to the frequency domain using the Fast Fourier Transform (FFT). The narrowband spectra were processed by the triangular melscale filterbank which can be expressed as:

$$H_m(k) = \begin{cases} 0 & f_k < f(m-1) \\ \frac{(f_k - f(m-1))}{(f(m) - f(m-1))} & f(m-1) \leq f_k \leq f(m) \\ \frac{(f(m+1) - f_k)}{(f(m+1) - f(m))} & f(m) < f_k \leq f(m+1) \end{cases} \tag{5.2}$$

where $f(\cdot)$ is the list of mel linearly spaced frequencies, m is the filter number, f_k is the frequency at FFT bin k. The mel filter bank was constructed such that there are 13 linearly spaced and 27 logarithmically spaced filters spanning the $13 - 6854$ Hz frequency range. MFCCs were then derived by computing the log of the mel-filtered spectrum and applying the discrete cosine transform (DCT).

5.2.4 Gammatone Frequency Cepstrum Coefficients (GFCCs)

GFCCs are mainly used in computational auditory sense analysis (CASA) studies to transform signals into the time-frequency (T-F) domain. The Gammatone filterbank has a better performance in modeling the auditory filterbank than the mel filterbank; hence, cepstral coefficients extracted using the Gammatone filter bank have a better speech recognition performance than MFCCs [9]. In GFCC, the equivalent rectangular bandwidth (ERB) scale is used. The impulse response of the Gammatone filter is given by,

$$g(t) = \frac{at^{n-1}\cos\left(2\pi f_c t + \varphi\right)}{e^{2\pi bt}} \tag{5.3}$$

where f_c is the filter center frequency, ϕ is the phase of the carrier, a is the amplitude, n is the filter order, b is the filter bandwidth, and t is the time in seconds. The value of the filter order and carrier phase were set to be $n = 4$, $b = 1.019$ ERB, $\phi = 0$. Laplace transform and the impulse invariance method were applied to transform the impulse response of the Gammatone filter into a discrete time equivalent filter. Extracted digital filters were applied to the framed speech to extract the energies of the filter banks. GFCCs were then computed by applying DCT to the log of gammatone filtered spectra.

5.3 Deep Learning

Deep learning is a machine learning method that uses a cascade of multiple layers for feature extraction and transformation. We considered the DNN architecture for the deep learning which consists of an artificial neural network (ANN) with multiple hidden layers between the input and output layers. The neurons of the DNN layers are trained (learned) using the input data only and therefore, do not require any human intervention. In this work, we used the hyperbolic tangent function (also called the symmetric sigmoid function) as the the neural transfer function for all the neurons of the hidden layers. It is explicitly defined by

$$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \in (-1, 1). \tag{5.4}$$

For the last output layer, we applied the pure linear transfer function. After calculating the error function, backward propagation was applied through the neural network from the output layer to the input layer to modify the network weights using the Levenberg-Marquardt optimization algorithm [75, 76]. The Levenberg-Marquardt algorithm is similar to quasi-Newton algorithms that guarantees an almost second-order training speed without having to compute the Hessian matrix. In fact the Hessian can be approximated as follows

$$H \simeq J^{\top} J \tag{5.5}$$

and the corresponding gradient vector is

$$g = J^{\top} e \tag{5.6}$$

where J is the Jacobian matrix that contains first derivatives of the network errors with respect to the weights and biases, and e is a vector of network errors. The Levenberg-Marquardt algorithm uses these approximations of the Hessian matrix and the gradient vector in the following Newton-like update rule:

$$x_{k+1} = x_k - (J^{\top} J + \mu I)^{-1} J^{\top} e \tag{5.7}$$

where μ is an adaptive tuning parameter such that for $\mu = 0$ the algorithm reduces to the Newton's method, and when μ is large it reduces to the gradient descent with a small step size. During the training, μ is decreased after each successful step (reduction in performance function) and is increased only when a tentative step would increase the performance function. In this way, the performance function is always reduced at each iteration of the algorithm.

5.4 Methodology

This research aims to utilize the DNN to generate a non-intrusive model that is used to predict TE voice quality. It is noted that deep learning training usually requires a large number of samples to be trained accordingly and predict the quality of a signal from a given set of input features. However, since generating a dataset of thousands of TE speech samples would require an immense time and effort and might be even

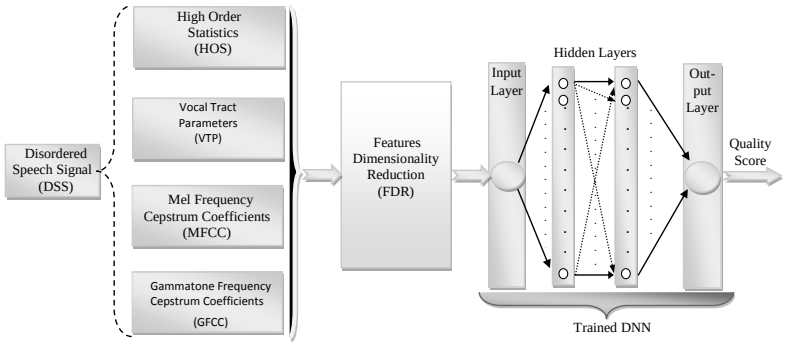

Figure 5.1: Proposed algorithm for disordered speech quality estimation. The DNN network is pre-trained using a custom synthetic dataset and then is used to map the extracted speech features (MFCC, GFCC, HOS, VTP) into a quality score.

impossible to find a sufficient number of samples, we adopted a different approach in this work. Indeed, our idea is to generate an *artificial dataset* of disordered speech signals from an existing normal speech dataset that is corrupted using speech-type noise and then scored using a validated objective quality index, *viz.* the Hearing Aid Speech Quality Index (HASQI) [48]. This generated dataset will be used to select the appropriate subset of features and to train the given DNN network. Once the features are selected and the DNN network is trained, the obtained algorithm can be used to assess the quality of the TE speech samples.

5.4.1 Artificial Subjective Dataset

First we considered a dataset of 53 clean (normal) speech signals. The sentence (from the Rainbow Passage[1]) is "*When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch*". All recordings

[1]The "Rainbow Passage" was developed by speech scientists and contains all of the sounds and sound combinations found in the English language.

were gathered in a sound-treated environment at sampling rate of 44.1 kHz with 16-bit quantization. Next, in order to create the perceptual attribute of roughness, the clean dataset is corrupted, at different SNR levels, by two types of speech-like noise: the MNRU [77] noise and the speech-shaped noise of [78], see Figure 5.2. The

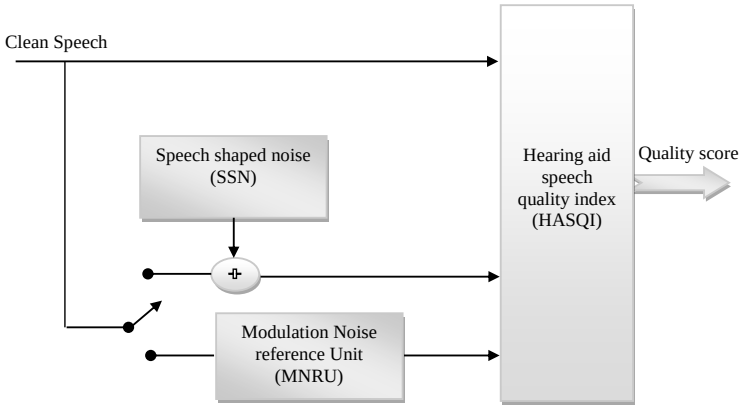

Figure 5.2: We have generated an artificial disordered speech dataset (2173 samples) using clean normal speech samples and speech like noise. The HASQI metric was then used to predict the perceived quality of each deliberately distorted speech signal. We compared between two types of speech like noise: speech shaped noise (SNN) and the Modulated Noise Reference Unit (MNRU) noise.

modulated noise reference unit (MNRU) noise is described in [77], and the output function is defined as follows:

$$y\left(i\right) = x\left(i\right)\left[1 + 10^{-Q/20}N\left(i\right)\right] \tag{5.8}$$

where $x\left(i\right)$ is the input signal, $y\left(i\right)$ is the output signal, $N\left(i\right)$ is white Gaussian noise, and Q is defined as the ratio, in dB, of speech power to modulated noise power. To generate different speech samples with varying degradations for each available

clean signal, we varied the value of Q between -5 and 35 dB resulting in 41 noisy speech samples corresponding to each clean speech sample. This yielded a total of $53 \times 41 = 2173$ disordered samples. For comparative purposes, and following the same procedure as discussed above, we also considered the generation of a second dataset using the additive colored noise reported in the literature [78]. More specifically, the noise specifications were set according to the long-term average speech spectrum data from Table 2 in [78] and will be referred to as speech shaped noise (SSN) throughout the paper.

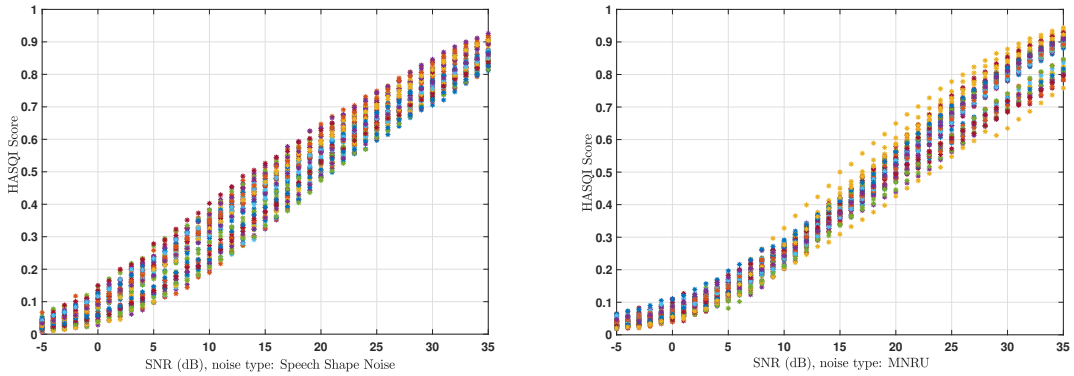

Figure 5.3: The HASQI [48] values for the two artificially generated speech datasets. At each SNR, the HASQI values were computed for each of the 53 normal speech samples.

The two generated databases were then evaluated using the HASQI and the obtained scores served as the target scores during the feature reduction and DNN training processes. It is pertinent to note here that the HASQI is the product of two indices: noise and nonlinear distortion index (called the nonlinear index), and linear filtering and spectral change index (called the linear index). The nonlinear index computes the time-frequency representations for both the original and degraded speech signals using a basic cochlear model. It combines a cepstral correlation term with

a vibration correlation term. The linear index is obtained by averaging the time-frequency representations across time, which quantifies the differences between the long-term average spectra (LTAS) of the test signal and the reference signal while ignoring the short-term differences in signal modulation and temporal fine structure [48]. Figure 5.3 depicts the distribution and range of HASQI values for the 53 speech samples at each SNR, for both types of corrupting noise. The statistics of the two synthetic datasets are given in Table 5.1. It is clear that we have chosen the SNR range to cover almost the whole spectrum of HASQI score which is between 0 and 1. This will allow to train a more accurate DNN network for the evaluation of the TE speech quality.

Dataset	Minimum	Maximum	Mean	Standard Deviation
Synthetic MNRU	0.019	0.944	0.407	0.281
Synthetic SSN	0.005	0.926	0.429	0.270

Table 5.1: Statistics of the HASQI scores for the two synthetic datasets.

5.4.2 Feature Selection and Reduction

Overfitting can be caused by a unreasonable higher dimensionality of the feature vector. In such situations, the extracted speech quality features must be reduced (in number) before being fed to the neural network to avoid overfitting. The aforementioned feature extraction methods were applied to the speech samples to estimate the speech characteristics. The extracted speech features included 60 GFCC features, 26 MFCC features, 54 VTP features, and 14 HOS features. First, forward stepwise regression (FSR) [68] is performed to prioritize the features within the group. Initially no predictors are included in the model. Then, at a first step, we check all

the possible models with one predictor against the coefficient of determination R^2 (R-squared) and the feature that gives a model with the highest R^2 is retained. The second step consists in checking all the models with two features by adding another feature to the previously selected feature. This procedure is repeated until we select all the available features. Note that the FSR algorithm stops also if, otherwise, the value of R2 reached 1 where in this case the remaining features are discarded. Finally,we obtain a natural ordering of the features by their importance. Then, we use K-folds cross validation method [69] to select the best set of features that guarantees the lowest prediction error (test error). This allows to avoid the problem of overfitting. For each number of selected features (obtained from the FSR), we use a 7-folds cross validation by training and testing support vector machine regression model with a linear kernel function. Figure 5.4 plot the out-of-sample mean square error (MSE) for each cross-validated model resulted from the selected features for the two synthetic datasets. From this figure we can determine the set of features that minimizes the out-of-sample MSE. In particular, we cut down the number of features from 154 features to 60 features consisting of 18 GFCC features, 16 MFCC features, 16 VTP features, and 10 HOS features by minimizing the out-of-sample mean square error (MSE) for the SMNRU synthetic dataset. Although, the synthetic SSN dataset minimizes the error at a lower number of features, we have chosen the same number of features for the future investigations to keep the comparison fair.

5.4.3 Deep Neural Network (DNN) Training

The DNN used in this work consisted of a 60 feature input layer, 2 hidden layers of 15 neurons each, and one neuron output layer that represented the perceived quality.

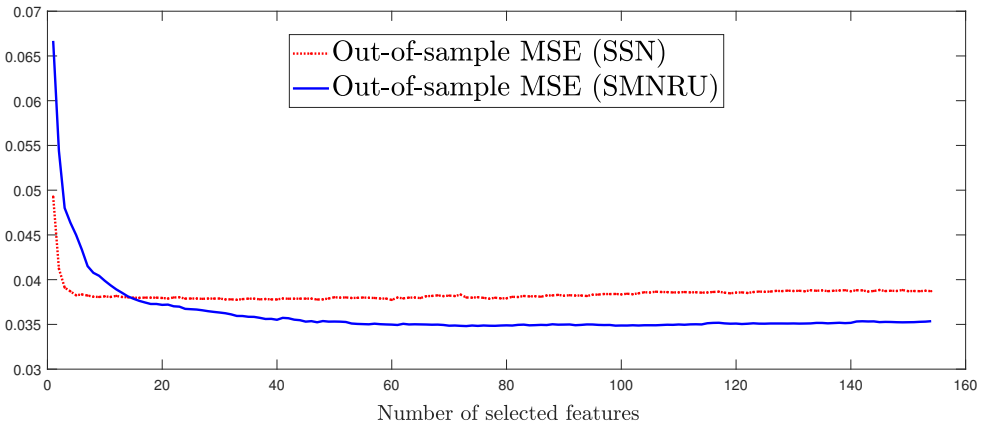

Figure 5.4: Out-of-sample mean square error during the feature reduction process. A total of 60 features is a reasonable choice that optimizes the out-of-sample MSE while guaranteeing the smallest number of features possible.

The Levenberg-Marquardt method was used as the optimization algorithm, sigmoid function was used as the activation function in the hidden layers and the tanh function was used at the output layer. The two noisy speech datasets (SSN and SMNRU synthetic datasets) were used for training the two diferent neural networks. Each dataset was randomly partitioned into three subsets: train, validation and test sets containing 70%, 15%, 15%, respectively, of each of the synthetic datasets. As the original normal speech database contained samples from 53 diferent speakers (both male and female), speaker dependency is not an issue and random partitioning was deemed appropriate. The prediction model training was iterated 60,000 times, and early stopping was enabled if the error on the validation dataset increases after reaching a minimum. Standard regularization, which modifies the performance function by adding a term containing the average sum of squares of network weights and biases, was applied. Dropouts were not utilized during the training phase. Fig.5.5 shows the plot of the true quality scores obtained by HASQI against the objective (predicted)

scores for all the samples of the datasets. The correlation values for all the datasets were about 0.99 (SMNRU) and 0.91 (SSN).

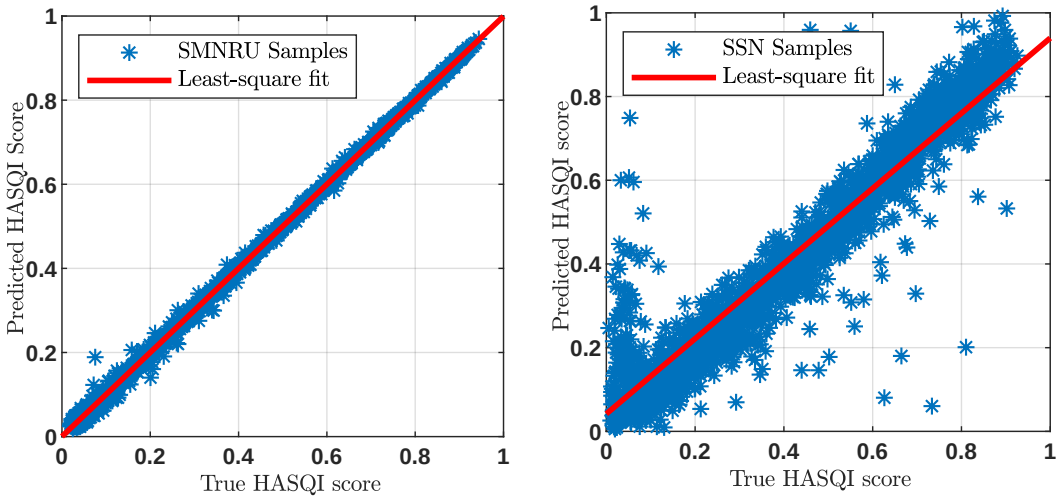

Figure 5.5: DNN training results for both artificially generated datasets.

5.5 Results

5.5.1 TE speech dataset

In this research, different TE speech databases are used to test the performance of the presented model. Four databases, denoted D_1, D_2, $D3$ and D_4 were collected and evaluated by researchers at the School of Communication Sciences and Disorders at Western University after obtaining ethics approval from the University's health sciences research ethics board. The speech samples were recorded from adult males between the ages of 45-65 years. All had undergone total laryngectomy and TE puncture voice restoration and all were at least one-year postsurgery at the time of their participation. All recordings were gathered in a sound-treated environment at a

sampling rate of 44.1kHz with 16-bit quantization. The second sentence of a standard reading (The Rainbow Passage), "The rainbow is a division of white light into many beautiful colors" was extracted from the full recording from all speakers and used for acoustic and perceptual measurements. For the auditory-perceptual phase of the study, the TE speaker samples were played back to a group of naive listeners who had no prior exposure to TE speech. The signals were played back in a randomized order and the listeners were instructed to rate the overall perceived severity/quality on a visual analog scale. The average of listener ratings were then used as the final subjective score.

5.5.2 Evaluation and Performance

The previously trained DNN model was used to estimate the ratings for these 4 databases of TE speaker recordings. The prediction model was iterated 60,000 times. To assess the performance of the proposed speech quality estimation algorithm as well as to compare with existing methods from the literature, four performance criteria were used [79]. First, the linear relationships between predicted quality scores and subjective ratings were quantified via the Pearson correlation, denoted ρ. Second, the ranking capability of the objective metrics was characterized by the Spearman rank correlation, denoted ρ_{spear}, which is computed in a manner similar to ρ but with the original data values replaced by their ranks. Next, a sigmoidal mapping function was used and once the objective values were mapped, a new Pearson correlation (termed ρ_{sig}) was computed and used as the third performance criteria [52]. The sigmoid

mapping is given by:

$$Y = \frac{\alpha_0}{1 + \exp(-(\alpha_1 X - \alpha_2))} \tag{5.9}$$

where α_0, α_1 and α_2 are the fitting parameters, X represents the predicted quality score, and Y the mapped predicted quality score. Lastly, the root square mean error, denoted RMSE, between the subjective and objective quality scores was used as our fourth performance criteria.

Tables 5.2, 5.3, 5.4 and 5.5 show the obtained results, for databases D_1, D_2, D_3 and D_4 respectively, when considering different objective speech quality algorithms. In particular, we have chosen to compare the performance of our algorithm (DNN-SSN and DNN-SMNRU speech modulation noise reference unit) against: voice breaks (VB), harmonics-to-noise-ratio (HNR), cepstrum peak prominence (CPP), smoothed cepstrum peak prominence (CPPs), telephony standard ITU-T P.563, energy capture ratio (ECR)[49], matching pursuit-based algorithm (MPAPESQ)[49] and matching pursuit-based algorithm (MPAHASQI). Figures 5.6-5.7-5.8-5.9 show the plots of the subjective ratings on the X-axis against the objective ratings on the Y-axis for the 4 TE speech databases. These results show that our proposed DNN-SMNRU algorithm has a superior performance in measuring the quality compared to the other objective metrics, including the DNN-SSN algorithm. The obtained correlations ranged between $0.72 - 0.8$ for SMNRU based DNN and $0.45 - 0.55$ for the SSN based DNN. This suggests that the MNRU is an effective noise generation scheme to mimic TE speech signal quality and could be used to generate artificial datasets for training more complex DNN architectures.

Metric	ρ	ρ_{spear}	ρ_{sig}	RMSE
VB	0.24	0.23	0.33	0.30
HNR	-0.10	-0.01	0.15	0.55
CPP	-0.26	-0.28	0.27	0.50
CPPs	-0.43	-0.41	-0.19	0.55
ITU-T P.563	0.21	0.27	0.19	0.56
ECR	-0.42	-0.53	0.61	0.23
MPA-PESQ	-0.77	-0.77	0.79	0.22
DNN-SMNRU	-0.72	-0.72	-0.71	0.20
DNN-SSN	-0.55	-0.56	0.55	0.34

Table 5.2: Correlation values for different objective metrics for database D_1.

Metric	ρ	ρ_{spear}	ρ_{sig}	RMSE
VB	0.16	0.19	0.18	0.34
HNR	0.40	0.33	0.36	0.56
CPP	0.30	0.31	0.30	0.51
CPPs	0.40	0.42	0.38	0.57
ITU-T P.563	0.50	0.44	0.51	0.56
ECR	0.57	0.62	57	0.24
MPA-PESQ	0.21	0.16	0.23	0.59
MPA-HASQI	0.36	0.44	0.42	0.25
DNN-SMNRU	0.75	0.74	0.76	0.24
DNN-SSN	0.55	0.53	0.62	0.26

Table 5.3: Correlation values for different objective metrics for database D_2.

5.6 Discussion

Since the TE speech production model differs from the normal speech production process, we have attempted to synthesize artificial TE speech datasets through the application of the MNRU type noise (standard ITU-T P.810) with the aim of generating samples that have similar “rough" and “noisy" perception as TE speech sentences. It is noted that the deployment of the standard ITU-T P.810 to modulate with clean speech samples led to the formation of a database that is similar in terms of per-

Metric	ρ	ρ_{spear}	ρ_{sig}	RMSE
VB	0.37	0.27	0.40	0.50
HNR	-0.51	-0.50	-0.10	0.51
CPP	-0.10	-0.14	-0.13	0.50
CPPs	-0.40	-0.21	-0.14	0.51
ITU-T P.563	-0.26	-0.22	-0.32	0.50
ECR	-0.47	-0.44	-0.47	0.50
MPA-PESQ	-0.47	-0.40	-0.40	0.48
DNN-SMNRU	-0.79	-0.78	-0.81	0.36
DNN-SSN	-0.68	-0.63	-0.69	0.39

Table 5.4: Correlation values for different objective metrics for database D_3.

Metric	ρ	ρ_{spear}	ρ_{sig}	RMSE
VB	0.23	0.27	0.0.15	0.37
HNR	0.10	0.01	0.01	0.51
CPP	0.54	0.52	0.51	0.30
CPPs	0.02	0.12	0.02	0.43
ITU-T P.563	0.02	0.12	0.02	0.43
ECR	0.46	0.40	0.46	0.33
MPA-PESQ	0.86	0.75	87	0.20
MPA-HASQI	0.75	0.65	0.80	0.20
DNN-SMNRU	0.73	72	0.74	0.20
DNN-SSN	0.43	0.37	0.44	0.35

Table 5.5: Correlation values for different objective metrics for database D_4.

formance to the TE speech. That was clear based on the obtained high correlation between the subjective and objective scores of the presented metric when applied to real (experimental) TE speech datasets. In fact, by considering different levels of MNRU noise, we were able to simulate different levels of TE speech quality without explicitly deriving the actual TE speech production model, one that may be complex and difficult to implement. The use of MNRU type noise to mimic the TE speech distortion gave better correlation compared to the use of the additive SNN type noise. The MNRU noise is multiplicative and provides a distortion that is perceptually sim-

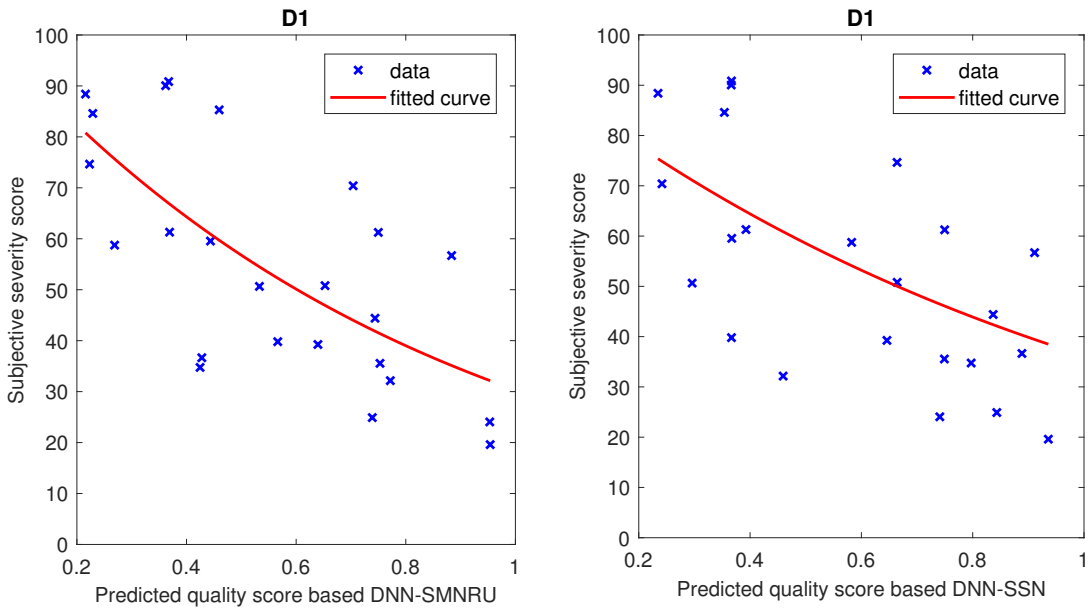

Figure 5.6: Subjective severity score for database D_1 and predictive quality score based on SMNRU and SSN.

ilar to TE speech.

Besides, we have utilized the deep learning algorithm to obtain the objective scores from the acoustic features of the TE speech samples. The neural network consisted of only two hidden layers to avoid the effect of overfitting and, thus, enhance the predictability of the presented objective metric. The presented model can be enhanced and increase the depth of the neural network by collecting more TE speech samples that can be used to directly train the neural network. The deep neural networks have the advantage that they do not need a set of features to be applied as the neural network can be applied to the speech samples directly. In order to be able to apply the neural networks to the TE speech, a larger number of speech

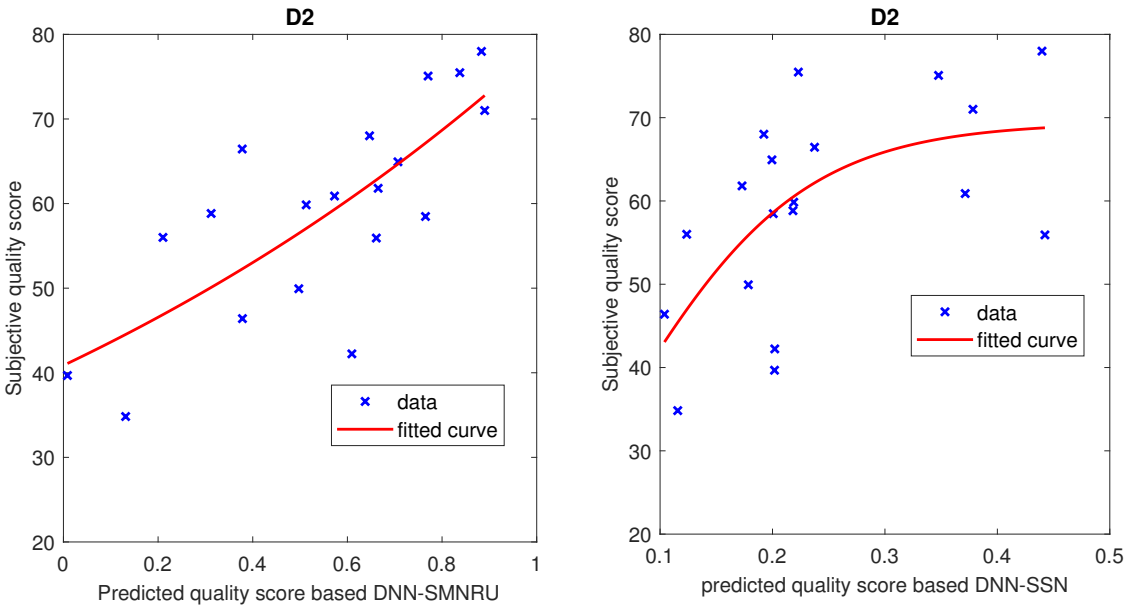

Figure 5.7: Subjective severity score for database D_2 and predictive quality score based on SMNRU and SSN.

samples will need to be collected to train the deep neural network. It is worth mentioning that Adam and SGD optimizers were investigated, but these optimizers did not show a statistical difference in performance from the Lavenberg-Marquardt optmizer. Lavenberg-Marquardt optimizer was preferred over other optimizers due to its better performance with small datasets such as the datasets presented in this researcher.

The proposed MNRU-based disordered speech synthesis approach, although correlates well with the subjective scores of the TE datasets at hand, can be improved if a more accurate TE speech synthesis model is derived. Future work will be carried out to investigate different disordered speech synthesis algorithms in order to synthe-

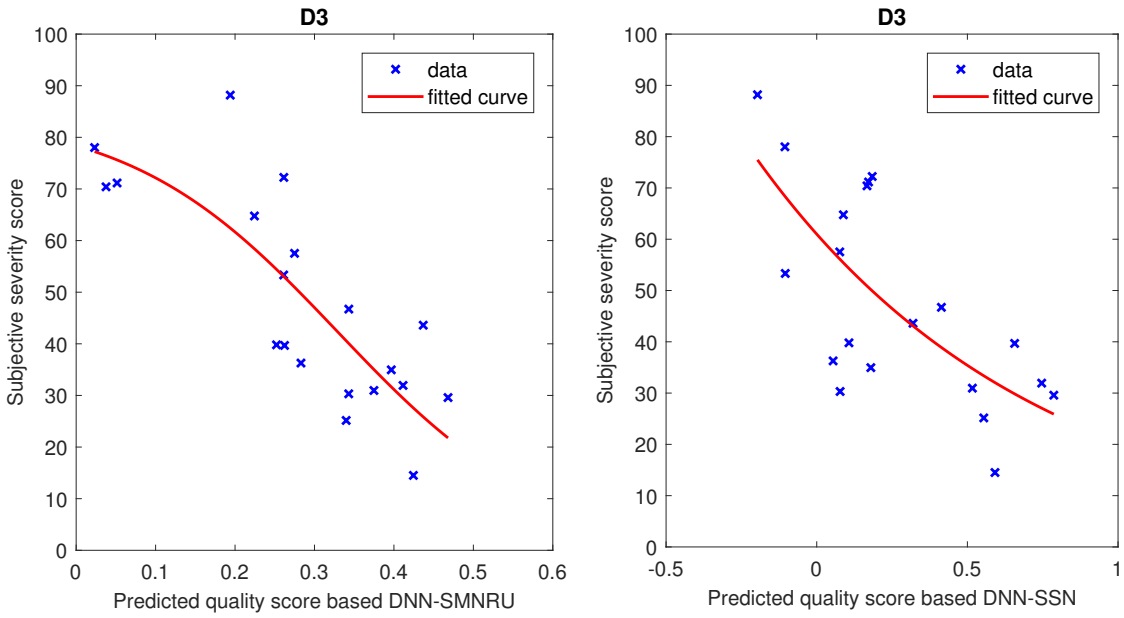

Figure 5.8: Subjective quality score for database D_3 and predictive quality score based on SMNRU and SSN.

sis disordered speech signals that are as close as possible to a real TE speech signal. This would pave the way to train larger and more accurate DNN networks for TE speech quality estimation.

It is noted that the combined model includes 18 features from GFCC and 16 features from MFCC despite the fact that GFCC and MFCC perform the same function of mimicking normal cochlear performance. In order to investigate the difference between the MFCC and GFCC performance, each feature was normalized according to the equation:

$$\hat{X} = \frac{X(i) - m}{N}. \tag{5.10}$$

where $X(i)$ is the number of the feature, m is the mean of the total number of

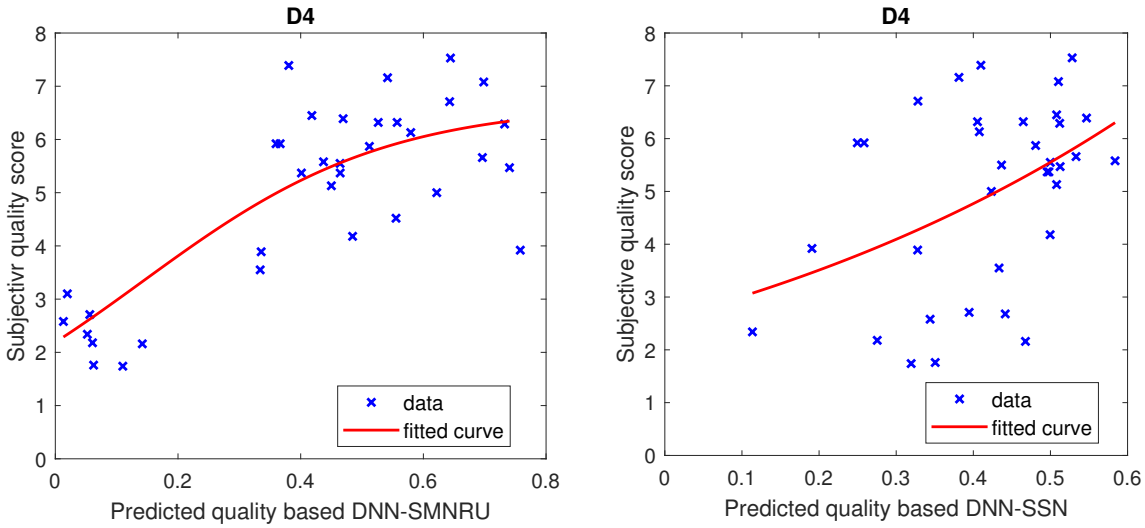

Figure 5.9: Subjective severity score for database D_4 and predictive quality score based on SMNRU and SSN.

features, N is the total number of features which is 26 in the case of MFCC and 60 in the case of GFCC, and $\hat{X}$ is the new number of the feature.

Fig 5.10 shows the scatter plot of normalized MFCC and normalized GFCC. This figure is to show why it is necessary to incorporate both of the MFCC and GFCC coefficients in the combined model. The figure shows that MFCC and GFCC figures are distributed across the spectrum, and MFCC coefficients give information about the high frequency components of the speech signal. On the other hand, GFCC coefficients have higher representation at low frequencies. The vertical line separates between the filter energies coefficients on the left and the delta coefficients on the right. It is noted that for lower frequency bands, GFCCs were selected to be more representing the quality than the MFCCs, with the latter more concentrated at the

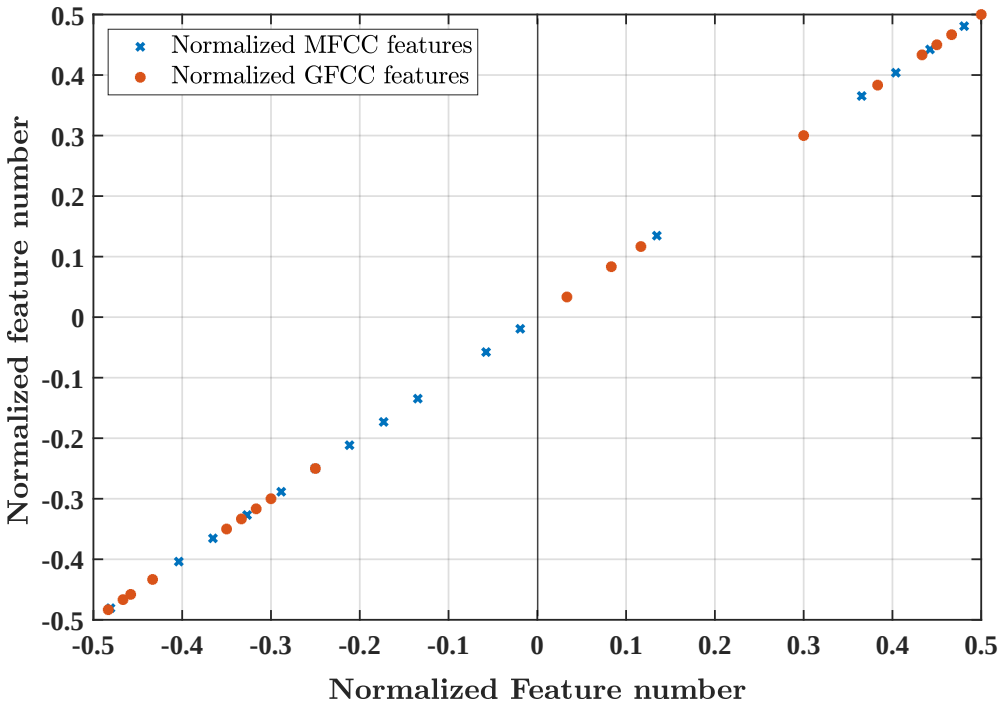

Figure 5.10: Normalized MFCC and GFCC coefficients.

high frequencies.The energy coefficients are the averaged energies in the filter bands across all the frames, while the delta coefficients are the first derivatives of the energy coefficients. For the delta coefficients, the MFCC delta coefficients were not selected at low frequencies. There were MFCC and GFCC selected delta coefficients at the high frequencies. It is clear from the figure that the delta coefficients at the mid band do not have an impact on the quality of the TE speech.

As mentioned earlier in the introduction, the PESQ metric has a high correlation value with TE speech subjective quality scores. However, this metric has a high computational cost, and the signal has to be processed in a complex way to estimate the objective quality.

5.7 Conclusion

The objective of this project aimed to develop a non-intrusive objective method to evaluate the quality of TE speech using DNNs as a supervised learning method. Two databases of 2173 training samples each were developed by modulating a dataset that contains 53 clean speech samples with 35 noise signals (at different SNR levels) that were generated following the standard ITU-T P.810 and the Speech Shape Noise approach to mimic the TE speech noise. These training datasets were evaluated by HASQI, and then these obtained scores served as targets for the training network. The input features extracted from the TE speech database included features of GFCC, MFCC, VTP, and HOS. The size of the feature input vector was reduced from 154 features to 60 features according to their correlation with the subjective scores. This trained DNN network had about 0.99 correlation value for the training, cross validation, and test databases. When the obtained DNN-based metric was applied to the experimental TE speech databases, it was found that the correlation between the subjective and objective scores correlation averaged value was about 0.75 for our four considered TE speech databases. This correlation value is higher than other objective measurements that are currently used or investigated in the previous chapter. Moreover, compared to the metrics derived in the previous chapters, this DNN-based algorithms shows a strong robustness property since it is validated on the four TE speech datasets.

Chapter 6

Conclusion

6.1 Summary

The purpose of this thesis was to develop a speech quality estimator tailored for the evaluation of the quality level of TE voice recordings. This objective is highly recommended in clinical applications to aid with the development of treatments for TE patients.

The first single-ended algorithm developed in this thesis was based on existing double-ended speech quality estimators such as the PESQ and the HASQI used in telecommunication and hearing-aids fields, respectively. Our idea was to generate an artificial reference signal using the matching pursuit algorithm which allowed to remove the incoherent parts from the disordered speech signal. Although this proposed approach was promising and correlates very well with some TE databases, it did not correlate sufficiently well with other TE databases. This can be explained by the fact that some TE speech distortions are difficult to separate from the coher-

ent speech parts using the atomic decomposition of the matching pursuit algorithm. Another plausible explanation is the fact that errors might be due to the PESQ and HASQI algorithms which were tuned and optimized for other applications. Moreover, the obtained MP-PESQ or MP-HASQI algorithm was relatively complex in terms of computational cost.

In an attempt to reduce the computational cost, we have investigated the use of statistical features extracted using low complexity algorithms such as the linear prediction analysis. The extracted features were trained using different regression models and showed promising results when correlated with subjective databases of TE speech. However, despite the simplicity of this type of algorithms, we were not able to obtain robustness across the different TE databases.

Next, we have studied the use of the deep neural networks to train model capable of predicting the TE speech quality. Since training these DNN models would require large datasets, we have opted for the generation of a large artificial database of noisy speech samples. This database was scored using the HASQI algorithm and then used to train the DNN. The obtained DNN was then used to predict the quality of the TE speech across different databases and the obtained results suggest that this DNN-based algorithm is robust and performs well for the TE speech quality evaluation.

6.2 Future Directions

Based on the study presented in this work, a number of recommendations exist for future work:

- This work focused on the objective assessment of quality of TE speech. Al-

though intelligibility and quality are two correlated attributes, a future study should focus on the objective assessment of the intelligibility of TE speech.

- Future researches should collect a large dataset of TE speech. This will lead to higher accuracy in the estimation of TE speech quality.

- Future work should be investigated to re-tune the parameters HASQI or PESQ algorithms in order to improve the robustness of the MPA-PESQ and MPA-PESQ algorithms. Also, further investigations on the role of the selected matching pursuit dictionary should be carried out across different TE speech datasets.

- The size of the DNN (number of layers) was limited due to the limitation of the dataset's size. Increasing the size of the training datasets will lead to implementing a deeper neural network and it will allow estimate the quality directly from the running speech samples.

Appendix A

CAPE-V

Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)

Name:_____ **Date:**_____

The following parameters of voice quality will be rated upon completion of the following tasks:

1. Sustained vowels, /a/ and /i/ for 3-5 seconds duration each.
2. Sentence production:
   a. The blue spot is on the key again.
   b. How hard did he hit him?
   c. We were away a year ago.
   d. We eat eggs every Easter.
   e. My mama makes lemon muffins.
   f. Peter will keep at the peak.
3. Spontaneous speech in response to: "Tell me about your voice problem." or "Tell me how your voice is functioning."

Legend: C = Consistent I = Intermittent
MI = Mildly Deviant
MO =Moderately Deviant
SE = Severely Deviant

			SCORE
Overall Severity	_____ MI MO SE	C I	/100
Roughness	_____ MI MO SE	C I	/100
Breathiness	_____ MI MO SE	C I	/100
Strain	_____ MI MO SE	C I	/100
Pitch	(Indicate the nature of the abnormality): _____ _____ MI MO SE	C I	/100
Loudness	(Indicate the nature of the abnormality): _____ _____ MI MO SE	C I	/100
_____	_____ MI MO SE	C I	/100
_____	_____ MI MO SE	C I	/100

COMMENTS ABOUT RESONANCE: NORMAL OTHER (Provide description):_____
_____

ADDITIONAL FEATURES (for example, diplophonia, fry, falsetto, asthenia, aphonia, pitch instability, tremor, wet/gurgly, or other relevant terms):

Clinician:_____

Bibliography

[1] Speech Language & Audiology Canada, "Communication health and aging," 2015, accessed on March 13, 2016. [Online]. Available: http://sac-oac.ca/sites/default/files/resources/communication_health_and_aging_brochure_web_en.pdf

[2] F. Cossa, "Measuring outcomes in speech-language pathology," *European Journal of Physical and Rehabilitation Medicine*, vol. 35, no. 2, p. 108, 1999.

[3] D. D. Mehta and R. E. Hillman, "Voice assessment: updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods," *Current Opinion in Otolaryngology & Head and Neck Surgery*, vol. 16, no. 3, p. 211, 2008.

[4] E. A. Peterson, "Toward validation of an acoustic index of dysphonia severity," Ph.D. dissertation, The University of Utah, 2012.

[5] V. Parsa and D. G. Jamieson, "Acoustic discrimination of pathological voice-sustained vowels versus continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 2, pp. 327–339, 2001.

[6] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, I. M. Moroz *et al.*, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *BioMedical Engineering OnLine*, vol. 6, no. 1, p. 23, 2007.

[7] D. Michaelis, M. Fröhlich, and H. W. Strube, "Selection and combination of acoustic features for the description of pathologic voices," *The Journal of the Acoustical Society of America*, vol. 103, no. 3, pp. 1628–1639, 1998.

[8] W. J. Levelt, *Speaking: From intention to articulation*. MIT press, 1993, vol. 1.

[9] C. M. Brown and P. Hagoort, *The neurocognition of language*. Oxford University Press, 1999.

[10] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[11] D. O'shaughnessy, *Speech communication: human and machine*. Universities press, 1987.

[12] A. G. Adami, "Automatic speech recognition: From the beginning to the portuguese language," in *9th International Conference on Computacional Processing of the Portuguese Language*, 2010.

[13] M. I. Singer and E. D. Blom, "An endoscopic technique for restoration of voice after laryngectomy," *Annals of Otology, Rhinology & Laryngology*, vol. 89, no. 6, pp. 529–533, 1980.

[14] J. Robbins, H. B. Fisher, E. C. Blom, and M. I. Singer, "A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production," *Journal of Speech and Hearing disorders*, vol. 49, no. 2, pp. 202–210, 1984.

[15] T. L. Eadie and P. C. Doyle, "Direct magnitude estimation and interval scaling of naturalness and severity in tracheoesophageal (te) speakers," *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 6, pp. 1088–1096, 2002.

[16] ——, "Scaling of voice pleasantness and acceptability in tracheoesophageal speakers," *Journal of voice*, vol. 19, no. 3, pp. 373–383, 2005.

[17] R. McDonald, "Objective evaluation of tracheoesophageal speech quality," Ph.D. dissertation, Faculty of Graduate Studies, University of Western Ontario, 2008.

[18] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice Hall, 1988.

[19] I. T. Union, "ITU-T recommendation P.800: Methods for subjective determination of transmission quality," *INTERNATIONAL TELECOMMUNICATION UNION*, 1996.

[20] J. Oates, "Auditory-perceptual evaluation of disordered voice quality," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 1, pp. 49–56, 2009.

[21] M. Hirano, *Clinical examination of voice*. Springer, 1981, vol. 5.

[22] G. B. Kempster, B. R. Gerratt, K. V. Abbott, J. Barkmeier-Kraemer, and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol," *American Journal of Speech-Language Pathology*, vol. 18, no. 2, pp. 124–132, 2009.

[23] M. P. Karnell, S. D. Melton, J. M. Childes, T. C. Coleman, S. A. Dailey, and H. T. Hoffman, "Reliability of clinician-based (grbas and cape-v) and patient-based (v-

rqol and ipvi) documentation of voice disorders," *Journal of Voice*, vol. 21, no. 5, pp. 576–590, 2007.

[24] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition." in *INTERSPEECH*, 2007, pp. 778–781.

[25] H. F. Wertzner, S. Schreiber, and L. Amaro, "Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders," *Brazilian journal of otorhinolaryngology*, vol. 71, no. 5, pp. 582–588, 2005.

[26] T. Leino, "Long-term average spectrum in screening of voice quality in speech: untrained male university students," *Journal of Voice*, vol. 23, no. 6, pp. 671–676, 2009.

[27] E. Mendoza, N. Valencia, J. Muñoz, and H. Trujillo, "Differences in voice quality between men and women: use of the long-term average spectrum (ltas)," *Journal of Voice*, vol. 10, no. 1, pp. 59–66, 1996.

[28] Y. Maryn, N. Roy, M. De Bodt, P. Van Cauwenberge, and P. Corthals, "Acoustic measurement of overall voice quality: A meta-analysisa)," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2619–2634, 2009.

[29] E. P.-M. Ma and A. L. Love, "Electroglottographic evaluation of age and gender effects during sustained phonation and connected speech," *Journal of voice*, vol. 24, no. 2, pp. 146–152, 2010.

[30] Y. Maryn and N. Roy, "Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity," *Jornal da Sociedade Brasileira de Fonoaudiologia*, vol. 24, no. 2, pp. 107–112, 2012.

[31] N. Roy, S. C. Mauszycki, R. M. Merrill, M. Gouse, and M. E. Smith, "Toward improved differential diagnosis of adductor spasmodic dysphonia and muscle tension dysphonia," *Folia Phoniatrica et Logopaedica*, vol. 59, no. 2, pp. 83–90, 2007.

[32] F. Klingholtz, "Acoustic recognition of voice disorders: A comparative study of running speech versus sustained vowels," *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2218–2224, 1990.

[33] Y. Qi, R. E. Hillman, and C. Milstein, "The estimation of signal-to-noise ratio in continuous speech for disordered voices," *The Journal of the Acoustical Society of America*, vol. 105, no. 4, pp. 2532–2535, 1999.

[34] F. Bettens, F. Grenez, and J. Schoentgen, "Estimation of vocal noise in running speech by means of bi-directional double linear prediction." in *INTERSPEECH*. Citeseer, 2003.

[35] L. Eskenazi, D. G. Childers, and D. M. Hicks, "Acoustic correlates of vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 2, pp. 298–306, 1990.

[36] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson, "Discrimination of pathological voices using a time-frequency approach," *Biomedical Engineering, IEEE Transactions on*, vol. 52, no. 3, pp. 421–430, 2005.

[37] S. Y. Lowell, R. T. Kelley, S. N. Awan, R. H. Colton, and N. H. Chan, "Spectral- and cepstral-based acoustic features of dysphonic, strained voice quality." *The Annals of otology, rhinology, and laryngology*, vol. 121, no. 8, pp. 539–548, 2012.

[38] S. Y. Lowell, R. H. Colton, R. T. Kelley, and Y. C. Hahn, "Spectral-and cepstral-based measures during continuous speech: capacity to distinguish dysphonia and consistency within a speaker," *Journal of Voice*, vol. 25, no. 5, pp. e223–e232, 2011.

[39] Y. D. Heman-Ackah, D. D. Michael, M. M. Baroody, R. Ostrowski, J. Hillenbrand, R. J. Heuer, M. Horman, and R. T. Sataloff, "Cepstral peak prominence: a more reliable measure of dysphonia," *Annals of Otology, Rhinology & Laryngology*, vol. 112, no. 4, pp. 324–333, 2003.

[40] C. R. Watts and S. N. Awan, "An examination of variations in the cepstral spectral index of dysphonia across a single breath group in connected speech," *Journal of Voice*, vol. 29, no. 1, pp. 26–34, 2015.

[41] S. N. Awan, N. Roy, M. E. Jetté, G. S. Meltzner, and R. E. Hillman, "Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: comparisons with auditory-perceptual judgements from the cape-v," *Clinical Linguistics & Phonetics*, vol. 24, no. 9, pp. 742–758, 2010.

[42] E. A. Peterson, N. Roy, S. N. Awan, R. M. Merrill, R. Banks, and K. Tanner, "Toward validation of the cepstral spectral index of dysphonia (csid) as an objective treatment outcomes measure," *Journal of Voice*, vol. 27, no. 4, pp. 401–410, 2013.

[43] K. Umapathy, S. Krishnan, V. Parsa, and D. Jamieson, "Time-frequency modeling and classification of pathological voices," in *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the*

Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint, vol. 1. IEEE, 2002, pp. 116–117.

[44] L. Malfait, J. Berger, and M. Kastner, "P. 563 — the ITU-T standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.

[45] R. McDonald, V. Parsa, and P. Doyle, "Prediction of the quality ratings of tracheoespohageal speech using adaptive time-frequency representations," in *Electrical and Computer Engineering, 2008. CCECE 2008. Canadian Conference on.* IEEE, 2008, pp. 001 715–001 718.

[46] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397–3415, 1993.

[47] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, 2001, pp. 749–752.

[48] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (hasqi)," *Journal of the Audio Engineering Society*, vol. 58, no. 5, pp. 363–381, 2010.

[49] Y. Ali, V. Parsa, P. Doyle, and S. Berkane, "Disordered speech quality estimation using the matching pursuit algorithm," in *The 30th Annual IEEE Canadian Conference On Electrical and Computer Engineering*, 2017.

[50] Y. S. E. Ali, V. Parsa, P. Doyle, and S. Berkane, "Disordered speech quality estimation using linear prediction," in *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, 2017, pp. 1–5.

[51] L. R. Rabiner and B. Gold, "Theory and application of digital signal processing," *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975. 777 p.*, 1975.

[52] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE signal processing magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[53] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248–248, 1961.

[54] R. ISO, "532 b: Acoustics-method for calculating loudness level," *International Organization for Standardization, Geneva*, 1975.

[55] A. A. Kressner, D. V. Anderson, and C. J. Rozell, "Evaluating the generalization of the hearing aid speech quality index (hasqi)," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 2, pp. 407–415, 2013.

[56] P. Kendrick, F. Li, B. Fazenda, I. Jackson, and T. Cox, "Perceived audio quality of sounds degraded by nonlinear distortions and single-ended assessment using hasqi," *Journal of the Audio Engineering Society*, vol. 63, no. 9, pp. 698–712, 2015.

[57] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (hasqi) version 2," *Journal of the Audio Engineering Society*, vol. 62, no. 3, pp. 99–117, 2014.

[58] P. Gray, M. Hollier, and R. Massara, "Non-intrusive speech-quality assessment using vocal-tract models," in *IEEE Proceedings on Vision, Image and Signal Processing*, vol. 147, no. 6, 2000, pp. 493–501.

[59] P. Gray, R. Massara, and M. Hollier, "Constraint-based pitch-cycle identification using a hybrid temporal/spectral method," in *Audio Engineering Society Convention 105*. Audio Engineering Society, 1998.

[60] A. Gaballah, V. Parsa, M. Andreetta, and S. Adams, "Objective and subjective speech quality assessment of amplification devices for patients with parkinson's disease," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 6, pp. 1226 –1235, 2019.

[61] D. Gabor, "Acoustical quanta and the theory of hearing," *Nature*, vol. 159, no. 4044, pp. 591–594, 1947.

[62] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.

[63] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 399–418, 1976.

[64] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1948–1956, Nov 2006.

[65] J. Lee and M. Hahn, "Automatic assessment of pathological voice quality using higher-order statistics in the lpc residual domain," *EURASIP Journal on Advances in Signal Processing*, 2009.

[66] J. B. Alonso, J. De Leon, I. Alonso, and M. A. Ferrer, "Automatic detection of pathologies in the voice by hos based parameters," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 275–284, 2001.

[67] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the lpc residual domain," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 217–231, 2001.

[68] R. M. Stolzenberg, "Multiple regression analysis," *Handbook of data analysis*, vol. 165, p. 208, 2004.

[69] R. R. Picard and R. D. Cook, "Cross-validation of regression models," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 575–583, 1984.

[70] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[71] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[72] L. Ljung, "System identification," in *Signal analysis and prediction.* Springer, 1998, pp. 163–173.

[73] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development.* Prentice hall PTR Upper Saddle River, 2001, vol. 1.

[74] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[75] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *Journal of the society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[76] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the marquardt algorithm," *IEEE transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, 1994.

[77] I. T. Union, "ITU-T recommendation p.810 (02/96), "Modulated noise reference unit (MNRU)"," *International Telecommunication Union*, 1996.

[78] D. Byrne, H. Dillon, K. Tran, S. Arlinger, K. Wilbraham, R. Cox, B. Hagerman, R. Hetu, J. Kei, C. Lui *et al.*, "An international comparison of long-term average speech spectra," *The journal of the acoustical society of America*, vol. 96, no. 4, pp. 2108–2120, 1994.

[79] S. Möller, W.-Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Wältermann, "Speech quality estimation: Models and trends," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, 2011.

Curriculum Vitae

Name: Yousef S. Ettomi Ali

Qualifications:

2004	UPM University: Master of Sciences in (Electronics Engineering). Peninsular, Malaysia
1991	College of Electronics Technology: Bachelor of Science in (Electronics Engineering). Bani-Walid, Libya.

Academic Work Experience:

1997 – August 2013	Lecturer: Al-Zawia College – Zawia, Libya.
2005 – 2007	Head of the Department Electrical & Electronics Department: Al-Zawia College – Zawia, Libya.
2004 – 2 0 0 5 .	Head of the scientific Committee / Al-Zawia College – Zawia, Libya.
1992 – 1997	Technical Engineer: Vocational Institute of Electrical and Mechanical professions – Zawia, Libya.

Publications:

- Microcontroller based Adjustable Closed-loop DC speed Controller (SCORD), IEEE, 2003.
- Microcontroller Performance speed control system, National power Conference (PECon), IEEE, 2004.
- Dynamic Simulation of the Speed-Torque controlled DC motor Drive Azzaytouna University Journal AUJ, 2014.

- Disordered Speech Quality Estimation Using Linear Prediction", in IEEE, PACRIM proceedings, pp. 1-5, 2017.
- Quasi-Reference Perceptual Speech Quality Estimation Using Matching Pursuit Algorithm", in IEEE, CCECE proceedings, pp. 842-846, 2017.

Secondary Experience:

Oct 2007 - Jul 2012	Adjunct lecturer: Al-Zawia University, Faculty of Engineering, Department Electrical & Electronics

Academic Responsibilities:

- *Curriculum development, exams preparations, and laboratories management.*
- *Teaching the following courses and setting up the related lab work.*

1. Power Electronics
2. Data Acquisition
3. Control Systems I
4. Control Systems II
5. Electronic Circuits I
6. Electronic Circuits II
7. Electric Circuits I
8. Electric Circuits I
9. Microcontroller
10. microprocessor
11. Electric Measurements
12. Electronic workshop
13. Software Engineering Math-lap Application

- **Supervised projects:**

Higher Diploma Project .	Robotic Design using Microcontroller (MC68hc11).
Higher Diploma Project	Automated irrigation using control system
Higher Diploma Project	Intelligent Traffic Light System
Higher Diploma Project	Alarm and Protection system

Higher Diploma Project	Design of control system for nursery room in hospital for new born Babies' using (PIC 16F84).
Higher Diploma Project	Design of remote-control system for home appliances using DTMF

Technical Work Experience:

1993 – 1997 Electronic Equipment Engineer: Al-Hana Electronics Company.

Responsibilities at Al-Hana Electronics Company:

Gadgets installation. Maintenance for Audio, TV, and Satellite systems.

Skills:

Software: MATLAB, MBLAB**, Multisim**

Good user for the following common test equipments such as Function generators, Oscilloscopes, Dc convertors, power meters, Avo meters.

Rewards:

2001 Libyan Higher Education Ministry Scholarship to do Master's Degree in Electronics, Malaysia.

2012 Libyan Higher Education Ministry Scholarship to do PhD in Electronics, Canada.

Languages:

- Arabic: native.

- English: Upper-Intermediate (Level C @ Culture-Works, London, Ontario).

Activities:

- Judo athlete, I participated in many national events.

- Voluntary worker for many charitable associations in Libya to help elderly in-need people.

References: Available on request

PhD student at UWO 2015

TA for these Subject

- ECE 2240A 2015
- ECE 4429 Fall 2016
- ECE 2241B 2017
- ECE 2240A 2017
- ECE 2274B2018
- ECE 2240A2018
- ECE 2274B2019