4-13-2018 1:15 PM

Learning-Based Reference-Free Speech Quality Assessment for Normal Hearing and Hearing Impaired Applications

Haniyeh Salehi
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

Recommended Citation

Salehi, Haniyeh, "Learning-Based Reference-Free Speech Quality Assessment for Normal Hearing and Hearing Impaired Applications" (2018). *Electronic Thesis and Dissertation Repository*. 5327.
https://ir.lib.uwo.ca/etd/5327

Abstract

Accurate speech quality measures are highly attractive and beneficial in the design, fine-tuning, and benchmarking of speech processing algorithms, devices, and communication systems. Switching from narrowband telecommunication to wideband telephony is a change within the telecommunication industry which provides users with better speech quality experience but introduces a number of challenges in speech processing. Noise is the most common distortion on audio signals and as a result there have been a lot of studies on developing high performance noise reduction algorithms. Assistive hearing devices are designed to decrease communication difficulties for people with loss of hearing. As the algorithms within these devices become more advanced, it becomes increasingly crucial to develop accurate and robust quality metrics to assess their performance. Objective speech quality measurements are more attractive compared to subjective assessments as they are cost-effective and subjective variability is eliminated. Although there has been extensive research on objective speech quality evaluation for narrowband speech, those methods are unsuitable for wideband telephony. In the case of hearing-impaired applications, objective quality assessment is challenging as it has to be capable of distinguishing between desired modifications which make signals audible and undesired artifacts. In this thesis a model is proposed that allows extracting two sets of features from the distorted signal only. This approach which is called reference-free (nonintrusive) assessment is attractive as it does not need access to the reference signal. Although this benefit makes nonintrusive assessments suitable for real-time applications, more features need to be extracted and smartly combined to provide comparable accuracy as intrusive metrics. Two feature vectors are proposed to extract information from distorted signals and their performance is examined in three studies. In the first study, both feature vectors are trained on various portions of a noise reduction database for normal hearing applications. In the second study, the same investigation is performed on two sets of databases acquired through several hearing aids. Third study examined the generalizability of the proposed metrics on benchmarking four wireless remote microphones in a variety of environmental conditions. Machine learning techniques are deployed for training the models in the three studies. The studies show that one of the feature sets is robust when trained on different portions of the data from different databases and it also provides good quality prediction accuracy for both normal hearing and hearing-impaired applications.

KEYWORDS: Speech quality evaluation, Intrusive assessment, Nonintrusive assessment, Wideband speech, Hearing aids, Wireless remote microphones, Auditory modeling, Support vector regression machine, Multivariate adaptive regression spline.

Acknowledgement

Firstly, I would like to express my sincere gratitude to my advisor Dr. Vijay Parsa for the continuous support of my PhD study and related research, for his motivation, patience, and immense knowledge. I am also thankful for all the learning opportunities he provided during the course of my study. I am grateful for all the hard questions which encouraged me to widen my research from various perspectives and taught me to be a critical thinker.

My sincere thanks goes to the applied DSP team in ON Semiconductor who provided me an opportunity to join their team. Working at ON Semiconductor was an excellent experience that helped me broaden my knowledge in hearing aid industry. Special thanks to Dr. Jim Ryan and Dr. Robert Brennan for their interest in my research and their insightful feedback. I am also grateful for the help and support I received from Adam Love during my work at ON Semiconductor.

I would like to thank my advisory committee, Dr. Susan Scollie and Dr. Hanif Ladak for their insightful comments and encouragement. I am grateful to all the members of the examining committee for accepting to review my thesis. Their comments further strengthened my thesis.

I am thankful to all the colleagues in National Centre for Audiology, whose support and encouragement made this journey easier.

Thanks to Dr. Alexandra Gerrie whose advice has made me who I am today. And last but not the least, I wish to thank my family and friends for their continued encouragement and support. Special thanks goes to my mother for her endless love.

Table of Contents

List of Tables

List of Figures

Nomenclature

AHD	Assistive Hearing Device
ANR	Adaptive Noise Reduction
ACR	Absolute Category Rating
BTE	Behind The Ear
BSD	Bark Spectral Distortion
BM	Basilar Membrane
CR	Compression Ratio
CC	Cepstrum Correlation
CIC	Completely In the Canal
CART	Classification and Regression Tree
DAI	Direct Audio Input
DSP	Digital Signal Processing
DFT	Discrete Fourier Transform
DSL	Desired Sensation Level
DMOS	Degradation Mean Opinion Score
DAM	Diagnostic Acceptability Measure
EC	Echo Cancellation
FFT	Fast Fourier Transform
FBE	Filter Bank Energy
FC	Feedback Canceler
FBE-HL	Filter Bank Energies incorporating Hearing Loss
GMM	Gaussian Mixture Model
HI	Hearing Impaired
HA	Hearing Aid
HL	Hearing Loss
HASQI	Hearing Aid Speech Quality Index

HATS	Head and Torso Simulator
HIRC	Hearing Instrument Review Committee
ITU	International Telecommunication Union
IS	Itakura-Saito
ISTS	International Speech Test Signal
IRS	Intermediate Reference System
IHC	Inner Hair Cell
LP	Linear Prediction
LPC	Linear Prediction Coefficient
LLR	Log-Likelihood Ration
LCQA	Low Complexity Quality Assessment
LCQA-HA	Low Complexity Quality Assessment for Hearing Aids
ModA	Modulation Spectrum Area
MWDRC	Multichannel Wide Dynamic Range Compression
MAD	Multiband Adaptive Directionality
MOS	Mean Opinion Score
MARS	Multivariate Adaptive Regression Spline
MUSHRA	MUltiple Stimulus test with Hidden Reference and Anchors
MPO	Maximum Power Output
NR	Noise Reduction
NH	Normal Hearing
NSIM	Neurogram Similarity Index Measure
NISA	Non-Intrusive Speech Assessment
OHC	Outer Hair Cell
PLS	Partial Least Squares
PCA	Principal Component Analysis
PLP-HL	Perceptual Linear Prediction incorporating Hearing Loss
PESQ	Perceptual Evaluation of Speech Quality
POLQA	Perceptual Objective Listening Quality Assessment
PEMO-Q	Perceptual Model-Quality Assessment
PEMOQ-HI	Perceptual Model-Quality Assessment for Hearing Impaired
PLP	Perceptual Linear Prediction
PSD	Power Spectral Density
PSE	Power Spectrum Envelope
RM	Remote Microphone
RMS	Root Mean Square
REAG	Real Ear Aided Gain

ρ	Pearson’s Correlation Coefficient
RBF	Radial Basis Function
σ	Standard Deviation of Error
SSN	Speech-Shaped Noise
SRT	Speech Reception Threshold
SVR	Support Vector Regression
SNR	Signal to Noise Ratio
SD	Spectral Distortion
SFM	Spectral Flatness Measure
SCF	Spectral Centroid Frequency
SRMR	Speech-to-Reverberation Ratio
SRMR-HA	Speech-to-Reverberation Ratio for Hearing Aids
SPL	Sound Pressure Level
SL	Sensation Level
VAD	Voice Activity Detection
ViSQOL	Virtual Speech Quality Objective Listener

Chapter 1

Introduction

1.1 Introduction

Speech intelligibility and quality metrics are highly desirable in the development and maintenance of diverse speech processing algorithms and communication systems. While speech intelligibility is related to speech understanding, speech quality refers to the overall listening experience. As noted in the literature (e.g. [2]), good speech intelligibility does not necessarily imply good speech quality. For example, in studying the relationship between speech intelligibility and quality, Preminger and Van Tasell [3] noted that intelligibility, pleasantness of tone, loudness, listening effort, and total impression are defined as five perceptual dimensions that play an important role in speech quality perception. They also stated that when intelligibility varies widely (i.e. from poor to excellent), so does the quality; but when all speech stimuli are of equal intelligibility, the other dimensions of speech quality influence subjective judgments [3]. Thus, a speech signal can be completely understood even when it is unnatural, unpleasant and has an inferior total impression. As such, speech quality assessment is important and plays a critical role in a number of applications such as telecommunications, broadcasting, Audiology, and Speech-Language Pathology. The focus of this thesis in on telecommunication and Audiology applications, which are described in detail next.

Telecommunication devices and systems change and get updated in order to better suit our needs and deliver more benefits. Wideband telephony has become very popular in the past years as it provides a better speech quality compared to narrowband telecommunication devices. Wideband telephony deploys a bandwidth of 0-8 kHz compared to 300-3.4

kHz in narrowband telephony for speech transmisstion. The benefits wideband telephony provides over narrowband telecommunication are widely studied and discussed in literature [4, 5, 6, 7, 8]. Echo and noise are two challenging issues in telecommunication devices. These challenges, however, become more significant in wideband telephony as the increased bandwidth allows more noise and echo to be transmitted [6]. High performance noise reduction (NR), and echo cancellation (EC) algorithms are two digital signal processing (DSP) modules deployed in such devices to overcome the aforementioned challenges [9, 10, 11, 12, 13, 14]. Benchmarking and fine-tuning the NR and EC algorithms have always been attractive to manufacturers, designers, and researchers [15, 16, 17]. Since delivering a better voice quality is the goal in communication applications, the assessments should be carried out with respect to perceived sound quality.

People with loss of hearing require assistive hearing devices (AHDs) to communicate with others. Hearing aids (HAs) and wireless remote microphones (RMs) are the most common treatment modality for listeners with mild to severe hearing loss (HL) [18]. It is shown that speech quality attributes such as clarity, fidelity, and naturalness are top drivers of satisfaction with and adoption of HAs [19]. As such, assessment and monitoring of AHD speech quality is of significant importance to designers, Audiologists, and researchers. Quality of speech in AHDs is not only affected by environmental influences like background noise and reverberation, but also by distortions caused by the DSP algorithms within these devices. Multichannel wide dynamic range compression (MWDRC), feedback canceller (FC), multiband adaptive directionality (MAD), and adaptive noise reduction (ANR) are the main modules within AHDs [20, 21, 22, 23, 24]. It is therefore important to quantify the impact of AHDs and their DSP algorithms on perceived speech quality under a variety of environmental operating conditions.

The purpose of this chapter is to: 1) review the major DSP algorithms deployed in normal hearing (NH) and hearing impaired (HI) communication devices, 2) introduce speech quality evaluation and the available methods, 3) layout the scope of this thesis, and finally 4) outline the thesis organization.

1.2 Telecommunication Applications

High frequency echo is one of the challenges in wideband telecommunication which is difficult to cancel. Research demonstrates that the human ear is more sensitive to high frequency echo [8]. To make it even worse, echo in the added frequency bands in wideband telecommunication is perceived louder [8]. Wideband noise is more annoying too because low frequency noise down to 50 Hz gets transmitted through wideband communication. The same scenario applies to noise in the high frequency portion. Signal enhancement algorithms such as NR and EC are deployed to overcome these challenges.

A low complexity NR algorithm for wideband speech coding is proposed in [25] while the challenges of delivering a good voice quality in wideband speech are covered in [6]. In [6, 26, 27] some of the limitations and challenges faced in echo cancellation is investigated. However, a robust speech quality prediction is desired to provide a systematic investigation.

1.3 Hearing Impaired Applications

The most common form of HL is a sensorineural HL which involves a multifaceted loss of hearing ability [28]. HI people do not hear all sounds and this causes difficulty in understanding speech because key parts of some phonemes are not audible. This deficit is referred to as decreased audibility [28]. Decreased dynamic range is another effect of HL on the auditory system. This means that the degree of increase in the threshold of hearing is much more than that of loudness discomfort [28]. Decreased frequency resolution and temporal resolution are the other two effects of HL on the auditory system where the former refers to the difficulty of separating sounds of different frequencies while the latter refers to the deficits in temporal envelope and fine structure analysis which leads to decreased ability in understanding speech in noise environments [28].

Current generation HAs employ DSP algorithms for delivering the appropriately amplified sound to the impaired ear to compensate for HL. A generic signal processing block diagram of a modern digital HA is shown in Figure 1.1. Typical DSP features in modern HAs include MWDRC, MAD, ANR, and FC, and the functionality of these features is described briefly below.

Hearing impairment involves a diminished sensitivity to lower level sounds, whereas perceiving loud sounds remains unchanged. As a result, there is a reduction in the range of detectable sound levels for HI individuals. WDRC algorithm compresses the range of detectable sounds by NH individuals into the detectable range for HI listeners. A gain is applied to low level sounds and this gain is reduced as the level of the sound increases. As a result, the higher level sounds are not amplified as much. This is shown in "Audibility

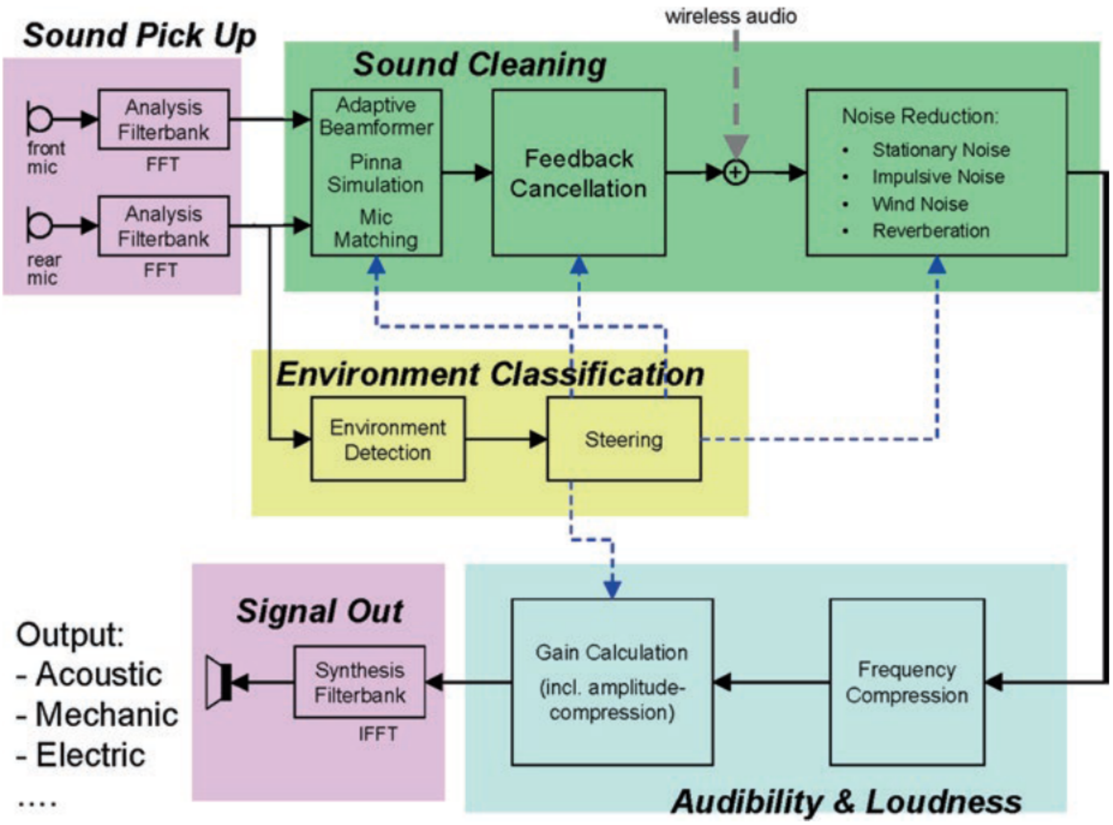

Figure 1.1: Generic signal processing block diagram of a modern digital HA [1]

and Loudness" in Figure 1.1. In a MWDRC, the same procedure is applied independently in different frequency regions or channels, based on HI person's audiogram [29, 30]. Souza [31] reviewed the literature that investigated the effects of WDRC on sound quality. It is concluded that automatic adjustment of HA gain with respect to input level is the core feature of compression which maintains speech audibility over a wide range of input levels. However, a reduced speech quality is reported by users when a large number of compression channels are used in conjunction with high compression ratios as they introduce artifacts to the HA output.

Perceiving sounds is harder in noisy environments than quiet environments. Directional microphones help in picking up the desired sound (typically speech) thereby reducing as much noise as possible and this helps in improving the signal-to-noise ratio (SNR). By combining two or more omnidirectional microphones a directional microphone is implemented. This is the first stage in the "Sound Cleaning" block in Figure 1.1. The elimination of noise or unwanted sounds is based on their angle of arrival at the HA's microphone. MAD is an algorithm that automatically selects the best polar plot in different frequency regions [32]. Effectiveness of directional microphones has been reviewed in [33]. It is concluded that directional microphones offer additional advantage compared to amplification alone. The advantage was optimized when a user-controlled switch was included and the subjects were trained for the environments where directional microphones provide better performance. In [34], user ratings for comfort and clarity are improved when directional HAs are used. Amlani et al. [35] reported better speech clarity scores for the directional microphone condition over the omnidirectional settings.

When desired and undesired sounds are spatially too close to be separated by directional microphones, the ANR algorithm is used (Figure 1.1). This algorithm is also appropriate for smaller HAs such as the completely in the canal (CIC) models, whose small size allows only a single microphone to be implemented [36, 37]. A typical ANR algorithm based on a spectral subtraction approach consists of a voice activity detection (VAD) block. This block controls a switch which is open when speech is detected and closed when speech is not detected. As a result, an estimate of noise spectrum is stored. This allows for the subtraction of the noise spectrum from speech plus noise magnitude spectrum. In modern HAs, this algorithm is implemented in a sub-band form. NR is evaluated through subjective measurements in [38]. A significant improvement in sound quality was reported when NR in a specific HA was enabled versus disabled. Learning abilities of children aged 11-12 are studied in [39]. The study suggests that NR may provide amplification with which children can tolerate and learn in noise. In some studies, however, it is shown that NR is not beneficial in certain situations where it is expected to see benefits ([40, 41, 42]).

In HAs, feedback occurs when a portion of HA output is captured by the HA microphone. Under favorable conditions, this feedback loop results in an annoyingly loud signal. Adaptive feedback cancellation algorithms estimate the transmission path between the HA speaker and microphone. The error between the desired and the actual output is taken and the adaptive processor adjusts its coefficients to minimize the error [43]. In [44], FC techniques for HAs are evaluated based on physical performance measures. In this study, the performance of a FC technique in four commercial HAs was assessed. In [45], the tradeoffs in adaptive FC for HAs were addressed. In this study, the parameters within three adaptive FC algorithms were adjusted in order to achieve better feedback cancellation and then the quality of speech was measured.

RMs represent a sub-class in the broader AHD category which when combined with the HAs, can help overcome listening difficulties [46]. It is well known that RMs perform better compared to HAs alone in acoustically challenging situations [47]. The reason is that they pick up the sound in a more optimal way. A transmitter microphone is used by the speaker and a receiver is used by the listener. The microphone is placed close to the talker's mouth, so the speech is picked up where less noise is added to the speech. The receiver transmits the sound to the listener's ears or directly to the HA if the listener is wearing a HA. The more noise there is, the closer the microphone must be to the sound source [48]. RMs come with their own DSP algorithms which could potentially affect the perceived quality. The previous studies [49, 50] investigated the speech intelligibility for RMs and further studies are required to look at their speech quality.

1.4 Speech Quality Evaluation

As explained above, both NH and HI applications deploy DSP algorithms to deliver good quality voice to users. These algorithms along with environmental distortions (e.g. noise and reverberation) affect quality of speech. As a result it is crucial to assess quality of delivered speech. This assessment helps in benchmarking different modules and algorithms within one device as well as the whole system under test. Moreover, in the AHD case, where the device needs to be fitted to patient's need, this assessment will be helpful to ensure the highest possible quality is delivered to the user.

Speech quality assessment is typically carried out through subjective or objective means. Although speech quality is a subjective measure in nature, subjective measurements are time- and resource-intensive. As such, an objective quality metric that correlates highly with subjective scores is attractive, as it can replace subjective measurements in determining the performance of different algorithms and devices. The two methods of quality assessment along with their advantages and disadvantages are discussed in the following sections.

1.4.1 Subjective Evaluation

There are several methods for assessing speech quality subjectively. These methods are generally classified into two groups: methods based on relative preference tasks and methods based on assigning a numerical value to the quality of the speech stimuli. Relative preference tasks include listening to clean speech and degraded speech signals and selecting the preferred degraded speech signal. In preference tests, listeners do not need to indicate the magnitude of their preference or the reason for their decision. Lastly, most preference tests produce a relative measure rather than producing an absolute measure. The rating tests, on the other hand, include listening to degraded speech signals and rating the quality of the speech signals on a numerical scale, typically a 5-point scale [51]. Following briefly describes these tests:

- Mean Opinion Scores (MOS) is the most common measure for user opinion. MOS is obtained by averaging the absolute category ratings (ACR). The comparison between the distorted signals with listeners' internal model of high quality speech is called the ACR. MOS is one of the methods recommended by IEEE subcommittee and the International Telecommunication Union (ITU) on subjective methods [51]. Another common measure is the degradation MOS (DMOS) tests. In DMOS, the subjects listen to both the original signal and distorted signal and rate the perceived degradation on a 5-point scale. DMOS tests are suitable for small signal degradations or impairments [51, 52].

- Diagnostic Acceptability Measure (DAM) evaluates several quality features of speech samples. There are 20 continuous rating scales, each dedicated to the evaluation of a given quality feature. There are three categories of scales: 1- features related to the speech signal (e.g. interrupted), 2- features related to the background noise (e.g. hissing, babbling), and 3- features covering both speech and background noise (e.g. intelligibility, acceptability) [53]. This is an expensive and time-consuming test because the listeners should be experienced.

To sum up, subjective tests are considered the "gold standard" in terms of evaluating speech quality. But, they are expensive and time-consuming. Subjective tests can be used only in the final stage of quality assessment and are not suitable for real-time applications.

1.4.2 Objective Evaluation

Objective speech quality measurements are based on physical measurements and mathematical models, which are generally calculated by comparing the original undistorted (clean) signal with the distorted signal. They do not require human listeners, so are less expensive and less time consuming. Since human listeners' judgement is not involved in objective evaluation, they give more consistent results compared to subjective evaluation.

In general, objective quality measures can be categorized based on two different criteria:

- The features which are used (parametric models vs. signal-based methods)
  Parametric models and signal-based models are different in the features they use for evaluation. Signal-based models use features extracted from either or both of the undistorted and distorted signals to predict the quality of the system under test. Parametric models use the system parameters and physical measures such as delay, echo and attenuation of the system under test.

- The information they need (intrusive vs. non-intrusive methods)
  Intrusive (full-reference) methods need both undistorted and distorted signals for evaluating the quality of speech. The reference signal should be sent to the algorithm under test and both reference and distorted signals are employed by the model to predict the speech quality. Non-intrusive or reference-free methods, on the other hand, work only based on the distorted speech. As shown in Figure 1.2, for intrusive speech quality estimation a feature set is extracted from both reference and degraded signals. Then the features are compared to estimate the amount of degradation, which is mapped to subjective scores to estimate the perceived speech quality. On the other hand, non-intrusive speech quality estimation is done by extracting features from the degraded signal only and directly mapping these features to the subjective scores. While intrusive methods are more powerful and accurate, there are applications where a reference signal is not available and intrusive methods are not applicable [54].

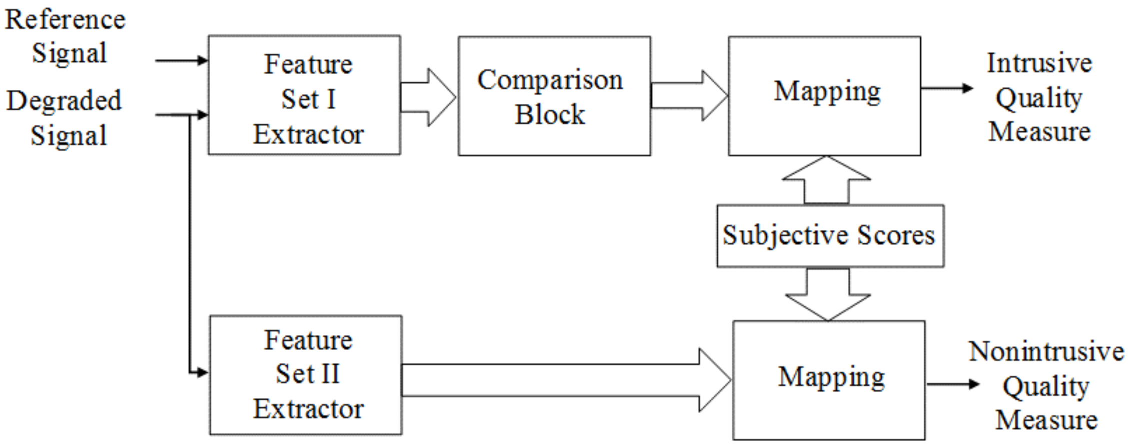

Figure 1.2: Intrusive assessment versus nonintrusive assessment

1.4.3 Figures of Merit for Objective Measures

Although there are different ways to measure speech quality objectively, what eventually matters is that it must be based on human perception. So, creating subjective quality databases is crucial in the assessment of objective measure success. As a result, the first criteria to assess the accuracy and strength of an objective measure is to determine if it follows the trends present in subjective scores (*ground truth*). The linear relationship is measured by Pearson's correlation coefficient between predicted quality scores and true quality score (MOS):

$$\rho = \frac{\sum_d (S_d - \bar{S_d})(O_d - \bar{O_d})}{(\sum_d (S_d - \bar{S_d})^2 \sum_d (O_d - \bar{O_d})^2)^{1/2}} \tag{1.1}$$

where, S_d are the subjective results and O_d are the objective results and $\bar{S_d}$ and $\bar{O_d}$ are their corresponding average values [55].

Another measure for assessing the agreement between two measurement methods is proposed by Altman and Bland (B&A) [56]. This measure is based on the quantification of the agreement between two quantitative measurements by studying the mean difference and constructing limits of agreement. In the B&A plot analysis a bias between the mean differences is evaluated and agreement interval is subsequently estimated. In B&A analysis, 95% of the differences of the second method compared to the first one fall within the agreement interval [56]. It should be noted that the B&A analysis is not typically performed for performance assessment of objective quality metrics.

There is an error associated with using the objective measure in place of the subjective measure. The standard deviation of the error is used as the second criteria and is given by:

$$\sigma_e = \sigma_d \sqrt{1 - \rho^2} \tag{1.2}$$

where σ_d is the standard deviation of S_d and ρ is the Pearson's correlation coefficient [55].

An objective metric that is good in predicting speech quality has a high correlation coefficient (close to 1) and small standard deviation of error (close to 0).

Steiger's Z-test [57] is another figure of merit used in this thesis. It is a test of the difference between two dependent calculations with one variable in common. Z-score is specifically used in psychological-related research where it is desirable to make statistical comparisons between correlation coefficients measured on the same individuals. The result of the calculation is called a Z-score. It can be compared in a 1-tailed or 2-tailed fashion to the unit normal distribution. If a 2-tailed test is performed, values greater than |1.96| are considered significant. The reader is referred to [57] for more details on this test.

1.5 Problem Statement and Thesis Scope

As discussed before, communication devices for NH and HI listeners are designed to transmit voice with high quality. In order to eliminate unwanted signals like background noise, echo, reverberation, and feedback, DSP modules are deployed. In HAs, amplification and compression are performed to compensate for loss of hearing. Since each individual has specific HL profile, HAs have to be fitted to meet user's needs. The aforementioned devices and systems, equipped with their unique, proprietary DSP features, may distort quality of speech received by NH and HI listeners. The distortion introduced by these algorithms are either speech distortion or noise distortion. Speech distortion is the type of artifacts that affect the speech signal itself, while noise distortion affects the background noise [58]. In addition to different types of distortions, NH and HI listeners perceive quality of speech differently.

For assessing the speech quality, objective measurements are more attractive than subjective measurements because of their advantages over subjective measurements. They are less expensive and less time consuming and they give more consistent results. A good objective quality measure should be able to pick up different distortions and predict the quality of speech for both groups of listeners. Within objective measurements, intrusive assessments are in general more accurate than nonintrusive ones. For example, Hearing Aid Speech Quality Index (HASQI) v.2 is an existing intrusive metric that has been investigated for both groups of listeners and different types of distortions [59]. However, an intrusive method is not applicable in practical situations requiring real-time quality monitoring. Moreover, intrusive methods require time alignment between the reference and distorted speech signals and frequency shaping of the reference signal based on the hearing loss profile. Existing nonintrusive methods work well for speech quality evaluation in wireless communications, voice over IP and telephony networks. The only non-intrusive algorithm for HAs is represented in [60]. This algorithm addresses three signal enhancement techniques and uses speech-to-reverberation modulation-ratio - HA (SRMR-HA) to predict speech quality. Since the SRMR-HA is only dependent on the relative distribution of modulation energy, its performance is affected in conditions where background noise has a "speech-like" modulation pattern. As a result, this method has a poor performance with multi-talker babble. Besides, it does not work well for traffic noise and its ability in predicting the performance of AHDs has not been investigated.

To my knowledge, no comprehensive study has looked into nonintrusive objective quality measures that predicts quality of speech for both NH and HI applications equally well. In this thesis, several quality measures are examined on NH and HI data sets and the results are reported. Two nonintrusive quality measures are proposed and tested on the data and the results are analyzed.

1.6 Thesis Organization

This thesis has been divided into 7 chapters as outlined as follows:

In Chapter 2, the structure of intrusive and nonintrusive quality measures is explained. The existing auditory modeling for NH and HI auditory systems is reviewed. The feature extraction block in the existing intrusive and nonintrusive measures is described. Three machine learning approaches commonly used for feature mapping are described.

In chapter 3, a review on HASQI as the existing intrusive metric for NH and HI applications with more emphasis on its HL model is done. Subsequently, two nonintrusive quality measures for NH and HI applications that incorporate the HL model in HASQI are introduced.

Chapter 4 outlines a study that focuses on validation of the proposed quality measures on NR algorithms in NH applications. Three other nonintrusive metrics and HASQI as an intrusive metric are applied to the same database. The nonintrusive measures that require learning are mapped to MOS using three learning methods. Based on a preliminary test, two out of three learning approaches are selected for further testing in which the training and testing procedure is performed on different portions of data to test the robustness of the models. It is followed by data analysis and discussion.

In Chapter 5, performance of the proposed quality measures is investigated on two HI databases. Multiple tests are performed where various portions of data is trained, and the model is tested on the remaining of the data. In the same manner as chapter 4, the results are compared with those of other nonintrusive measures and HASQI.

In Chapter 6, the details of a new database containing stimuli from four RM devices paired with a HA are presented. The procedure followed for collecting the stimuli is described and the relative performance of the RMs is benchmarked through objective, instrumental predictors of perceived speech quality by HI listeners. As the second study, the models trained on HI databases in Chapter 5 are tested on the stimuli for RM data and the performance of the models is compared with other nonintrusive metrics.

Chapter 7 covers an overall conclusion with suggestions for the future work.

Chapter 2

Speech Quality Evaluation

2.1 Introduction

In this chapter methodological aspects of objective quality evaluation is discussed. The process of perceiving quality in human's brain is complex and relatively unknown. However, it is clear that there has to be a comparison for quality judgment. Our brain is able to assess the quality of speech because of its previous experience with a wide range of audio qualities. Likewise, an objective quality measure has to have a brain too. Although this block has different names such as decision-making block, mapping block, cognitive model, machine learning, or fitting block, it takes care of mapping the feature set to a final quality index. Unlike our brain that perceives quality by listening to the whole audio signal, the learning block in an objective measure cannot process the whole signal. Hence, it is necessary to extract useful and relevant information from stimuli to derive a quality index. This is done in feature extraction block. In order to get a high accuracy out of an objective metric, it is crucial to extract features that contain relevant information for speech quality.

There are several factors contributing to the success of an objective model. Extracting features that have high correlations with subjective scores is one factor. Moreover, research shows that features extracted by auditory models correlate better with subjective scores compared to signal-based features [61, 62].

In this chapter, some of the objective speech quality metrics for NH and HI applications are reviewed. As per mapping block, partial least-squares (PLS) regression, support vector regression (SVR) machine, and multivariate adaptive regression spline (MARS) approaches are chosen to be reviewed.

2.2 Feature Extraction for Objective Speech Quality Measures

Existing literature suggests that objective quality measures use one or a combination of feature categories [63, 64, 65, 66].

Waveform-comparison measures are the simplest class of algorithms. They need low computational complexity, however they usually do not result in high correlations with subjective measurements. SNR measures are one of the oldest and most common waveform-comparison measures. It requires both clean and distorted signals. SNR can be calculated as follows:

$$SNR = 10\log_{10}\frac{\sum_{n=1}^{N} x^2(n)}{\sum_{n=1}^{N}\left(x(n)-\hat{x}(n)\right)^2} \tag{2.1}$$

where $x(n)$ is the clean speech, $\hat{x}(n)$ the distorted speech and N the number of samples [65]. Due to a wide range of distortions this definition does not correlate well with subjective quality scores. Other variations to the classical SNR are derived which show much higher correlation with subjective quality. In the classical SNR the portions of signal where speech energy is large and noise is inaudible are washed out by portions where speech energy is small and noise is high. So segmental SNR is introduced:

$$SNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Lm}^{Lm+L-1} x^2(n)}{\sum_{n=Lm}^{Lm+L-1}\left(x(n)-\hat{x}(n)\right)^2} \tag{2.2}$$

where L is the frame length and M is the number of frames in the signal. Using logarithm ensures that the frames with a large ratio are weighted less, while frames with a small ratio are weighted somewhat higher. The SNR_{seg} matches the perceptual quality well, however if the speech signal contains excessive silence, the overall SNR_{seg} will decrease because silent frames result in large negative . The upper and lower limits for SNR_{seg} are typically between 35 dB and -10 dB [15].

Frequency-domain techniques are widely used in speech quality prediction. They are shown to be more consistent with human perception and are not sensitive to time shifts

[64, 52]. Spectral distortion (SD) measure is one example in this category. An objective measure based on Bark spectral distortion is proposed in [67] for predicting quality of speech coders. BSD is the average squared Euclidean distance between spectral slopes between the original and coded speech signals.

Measures based on linear prediction (LP) and auditory-based or perceptually-motivated measures are the next two categories within objective quality measures. Perceptually-motivated measures are specifically very attractive as these measures incorporate knowledge of the human perceptual system and as a result correlate well with subjective scores. Normally, the number of features extracted in these two methods is reduced and then mapped to the subjective scores. Hence, these two steps are discussed first and then some of the existing LP-based and auditory-based objective methods are reviewed.

2.3 Dimensionality Reduction and Mapping Function

The main goals of the dimensionality reduction procedure are to retain the features that have a high degree of correlation with subjective data, and to minimize redundancy in predictive information carried by different features. Subsequently, the reduced feature set has to be mapped to a single predicted speech quality score for the speech sample under test. Mapping function involves finding the true regression function $f(x)$ that determines the relationship between the observations on the quality predictions, $y = (y_1, ..., y_n)'$, and the extracted features, $\mathbf{X} = (\mathbf{x_1}, ..., \mathbf{x_n})'$. The regression function has a form of:

$$y_i = f(x_i) + \varepsilon_i \tag{2.3}$$

where ε_i has normal probability distribution, $N(0, \sigma^2)$.

Averaging, Summation, and Minkowski summation are among the simple techniques that have been tried. Regression-based techniques are the most popular methods of training models [68]. In [69], dimensionality reduction is performed by principal component analysis (PCA). Then the reduced subset of features are mapped to speech quality by linear regression. Bayesian modeling has been utilized in [70, 71]. Gaussian mixture models (GMM) have also been exploited for feature mapping [72, 73, 69]. Another technique known as MARS [74] has also been explored in [55, 73, 75] for feature mapping. SVR [76] is a kernel-based technique which is shown to be powerful in many signal processing applications [77, 78, 79, 80] and speech quality prediction [81].

Every mapping technique has some benefits and drawbacks and its success mainly depends on the nature of extracted features and their relationship. So in this work three widely used techniques are studied. A two-step dimensionality reduction followed by linear regression is exploited as a simple mapping technique which works based on linear relationships between features. MARS and SVR are chosen as more powerful tools that account for nonlinear relationships between features.

Two-Step Dimensionality Reduction and Linear Regression

Dimensionality reduction is usually utilized to improve the performance of mapping. Through dimensionality reduction, the features that effectively carry various properties of the signal are retained. In the first step, a subset of features are selected through a correlation analysis. The correlation-based index is given by:

$$\varepsilon(i) = \left| \frac{\rho(i)}{\sum_{j=1}^{N} |\rho_{ij}|} \right| \tag{2.4}$$

where $\rho(i)$ is the Pearson correlation coefficient of feature i with averaged subjective quality ratings, and ρ_{ij} is the correlation coefficient of feature i with feature j, and N is the number of features in the global set before dimensionality reduction.

PCA is the best known linear feature extraction algorithm [82]. It is a non-parametric, unsupervised method of extracting relevant information from data sets. PCA linearly combines features to re-express the feature space as an orthogonal (uncorrelated) projections of high dimensional data [83].

A subset of the rank-ordered features were then transformed using the PCA and combined using a linear regression function.

Multivariate Adaptive Regression Spline (MARS)

MARS [74] is a data-driven approach, which adaptively derives basis functions from the feature set and linearly combines the basis functions to best match the subjective data. The MARS method employs forward selection and backward deletion procedures to retain features most important for quality prediction.

MARS builds a model as a product of spline basis functions. The number of functions and knot location are determined by data. MARS has a model of the form:

$$f(x) = a_0 + \sum_{K_m=1} f_i(x_i) + \sum_{K_m=2} f_{ij}(x_i, x_j) + \sum_{K_m=3} f_{ijk}(x_i, x_j, x_k) + ... \tag{2.5}$$

where a_0 is the coefficient of the constant basis function B1. The first sum is over all basis functions that involve only a single variable. The second sum is over all basis functions that involve two variables, representing two-variable interactions and the third sum represents the contributions from three-variable interactions and so on. The basis functions B_i are given by:

$$B_i(x) = \prod_{i=1}^{J_i} [s_{ij}(x_{w_{ij}} - t_{ij})]_+, \; i = 1, 2, \cdots, k \tag{2.6}$$

where $[.]_+ = max[0, .]$, J_i is the degree of interaction of basis function B_i, the s_{ij} are the sign indicators taking values ± 1, the t_{ij} are knot points and the w_{ij} give the index of the predictor variables which is being split on the t_{ij} [74].

Support Vector Regression (SVR) Machine

The goal in SVR [76] is to find the regression function $f(x)$ based on training data. In ε-SV, regression the goal is to find a function $f(x)$ that has at most ε deviation from the subjective quality scores, and at the same time is as flat as possible. $f(x)$ has a form of:

$$f(x) = \langle \omega, x \rangle + b \tag{2.7}$$

where X denotes the space of input features, $< ., . >$ denotes the dot product in X, ω is the weight vector, and b is the bias term. The goal is to find the weights and bias term such that the errors between predicted and true values are less than ε. Sometimes, however, it is not feasible to find the function $f(x)$ that satisfies such constraints. In such cases, we introduce slack variables ξ_i, ξ_i^*. Therefore, the optimization problem solves:

$$minimize \qquad \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{2.8}$$

$$subject \quad to \qquad \begin{cases} y_i - \langle \omega, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle \omega, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

where ξ_i is the upper training error, ξ_i^* is the lower training error and C is the penalty parameter of the error term. It is shown that the equation (2.7) can be written as:

$$f(x) = \langle \omega, x \rangle + b = \sum_{i=1}^{n_{sv}} (\eta_i^* - \eta_i) K(x_i, x) + b \tag{2.9}$$

where η_i^* and η_i are the Lagrange multipliers used in the Lagrange function optimization, n_{sv} is the number of support vectors, and $K(x_i, x)$ is the kernel function. In SVR, learning is based on support vector (i.e., critical points). It can be seen from (2.9) that the predicted value is a weighted sum of the distances between the support vectors and test vector [81].

The following sections cover some of the existing the LP-based and perceptually-motivated quality measures.

2.4 Measures Based on Linear Prediction

Measures based on LP of speech is another quality prediction technique. Speech production process can be modeled with an LP model. Some objective measures use the distance between two sets of linear prediction coefficients (LPCs) of the clean and distorted speech. The Log-Likelihood Ratio (LLR) measure is a distance measure that measures the difference between two speech signals and can be calculated as follows:

$$d_{LLR}(a_d, a_c) = \log \frac{a_d R_c a_d^T}{a_c R_c a_c^T} \tag{2.10}$$

where a_c is the LPC vector for the clean speech, a_d is the LPC vector for the distorted speech, a^T is the transpose of a, and R_c is the auto-correlation matrix for the clean speech [84].

The Itakura-Saito (IS) is another LPC-based measure that can be calculated as follows:

$$d_{IS}(a_d, a_c) = \left[\frac{\sigma_c^2}{\sigma_d^2}\right]\left[\frac{a_d R_c a_d^T}{a_c R_c a_c^T}\right] + \log\left(\frac{\sigma_c^2}{\sigma_d^2}\right) - 1 \tag{2.11}$$

where σ_c^2 and σ_d^2 are the all-pole gains for the clean and distorted speech [85].

A low-complexity, nonintrusive speech quality assessment (LCQA) for telephony networks is represented in [52]. This algorithm does not model the distortion explicitly. Six structural features are extracted from the signal. These features and their time derivatives are used to estimate speech quality. These six features are spectral flatness, spectral dynamics, spectral centroid, excitation variance, speech variance and pitch period. The spectral flatness measure (SFM) measures the shape of the power spectral density (PSD) [86]. This measure is related to the strength of the resonant structure in the power spectrum [52]. Spectral dynamics is used in speech coding and speech enhancement. It is shown that the dynamics of the power spectrum envelope (PSE) plays an important role in the perceived distortion [87]. Spectral dynamics is in fact a measure of how smooth a PSE is evolving. The next feature is spectral centroid which determines where most of the signal energy concentrates. This measure is associated with the brightness of a sound. Spectral centroid frequency (SCF) is the weighted average frequency in a sub-band. The weights are the normalized energy of each frequency component. This measure can detect the approximate location of formants (peaks) [88]. In LCQA, pitch period is used as a feature for quality assessment. Since pitch is important in distinguishing segmental categories in tonal languages, it can help in speech quality assessment. It is shown in [89] that the accuracy of this feature in noisy speech signals correlates with the accuracy of the objective quality evaluation of the speech signal.

The Importance-weighted Signal to Noise Ratio (*iSNR*) metric, which has been previously applied for prediction of speech intelligibility for normal hearing listeners [69, 90], is employed as:

$$iSNR(i) = 10 \times \sum_{k=1}^{N_f} I(k) \log_{10} \frac{max(0, P_x(i,k) - P_{\hat{v}}(i,k))}{P_{\hat{v}}(i,k)} \tag{2.12}$$

where $P_x(i,k)$ and $P_{\hat{v}}(i,k)$ are the input signal and estimated noise powers in the i^{th} frame and k^{th} frequency band respectively, $I(k)$ is the band importance function, and N_f is the number of frequency bands. The *iSNR* calculation is restricted to voiced portions of the speech signal, and therefore a suitable voice activity detection (VAD) algorithm is employed prior to the computation of frame *iSNR*.

LCQA-HA [75] is a nonintrusive metric for evaluating the speech quality of HAs. In LCQA-HA, spectral flatness, spectral dynamics, spectral centroid, excitation variance, and speech variance are the per-frame features extracted from the recordings. Pitch period is unaffected by HA processing [e.g., [91]] and as such was excluded from the feature set. Each speech recording was segmented into 20 ms non-overlapping frames and an 18th order LPC model was utilized. The individual features and their time derivatives were computed using the LPC model parameters, giving rise to a core set of 10 per-frame features. The global statistical properties of the per-frame features, viz. mean, variance, skewness, and kurtosis, across the entire recording resulted in a 40-dimensional feature vector. Unlike the original LCQA, no threshold-based frame selection was undertaken, while the dimensionality reduction was handled by the mapping functions, discussed later. In order to calculate the iSNR parameter, the frame-by-frame noise power spectrum was estimated using the minimum statistics algorithm. The frame-specific input and estimated noise power spectra were then grouped into $1/3^{rd}$ octave bands, and the frame iSNR is calculated using 2.12, with $I(k)$ representing the $1/3^{rd}$ octave band importance function [90]. Finally, the *iSNR* values of those frames identified by the VAD algorithm [90] were averaged and the averaged value appended to the above LCQA feature set. In ITU-T P.563 [15], the parameters based on higher order statistical characterization of cepstral and LPC coefficients are included in the objective quality prediction procedure. Since LPC coefficients are already calculated in LCQA, the averaged LPC skewness and LPC kurtosis parameters are included into the feature set as well.

Sharma et al. [92] introduced the non-intrusive speech quality assessment (NISA) which expanded the LCQA feature set by adding additional features such as the iSNR, and used the Classification and Regression Tree (CART) feature mapper.

2.5 Perceptually-Motivated Quality Measures

Perceptual Evaluation of Speech Quality (PESQ) was an international standard for estimating the MOS from both the clean signal and its degraded signal. The input and the degraded speech signals are converted into an internal representation using a perceptual model. These two signals are then time-aligned and the difference in the internal representation of the two signals is computed and used by the cognitive model to estimate the MOS.

PESQ was standardized by the ITU for narrowband telephony applications [93]. More recently, an extension of PESQ has been standardized as perceptual objective listening quality assessments (POLQA) [94]. POLQA is an intrusive metric that includes quality assessment of noise suppressed signals. It covers wider range of distortions and speech bandwidths. P.563 [15], is the nonintrusive equivalent of PESQ which is recommended by ITU-T for narrowband telephone networks and speech codecs. ITU-T P.563, extracts a total number of 51 speech features from signal. This measure is based on models of the human vocal tract and the human perception of distortions in a speech signal. First of all, the speech signal is preprocessed, which includes intermediate reference system (IRS) receive filtering, speech level adjustment and separating voice and non-voice parts by using a VAD. Then the distortion class is defined and speech parameters are extracted. The main distortion classes are unnaturalness of speech, basic speech quality, robotic voice, unnatural voice, strong additive noise, background noise, low segmental SNR, interruptions, mutes and clipping. After determining a dominant distortion class, mapping to final quality estimate is done.

Perceptual Model-Quality Assessment (PEMO-Q) [95], is another intrusive measure developed for predicting the speech quality of wideband speech samples degraded by audio codecs. The temporal and spectral masking in human auditory system is modeled in PEMO-Q. PEMOQ-HI [96] is an extension to PEMO-Q that models the HI auditory system. In this measure, outer hair cell (OHC) and inner hair cell (IHC) losses are modeled by instantaneous expansion and an attenuation stage before the adaptation stage.

Virtual Speech Quality Objective Listener (ViSQOL) is a full-reference metric which has been proposed in [97]. This metric has a particular focus on VoIP degradations and has been tested on a range of common background noises. In ViSQOL, both reference and degraded signals are passed through five major processing steps: pre-processing, time alignment, predicting warp, similarity comparison, and a mapping block. The auditory periphery model proposed in [98] is utilized to simulate the middle ear and inner ear. Neurogram similarity index measure (NSIM) [99] is used in ViSQOL to derive the quality index. NSIM is a distance measure that evaluates the auditory nerve discharges by comparing the neurogram for reference speech to the neurogram from degraded speech. Finally a sigmoid-like mapping function is used for deriving the final quality index [97].

HASQI v.1 [100] is an index which has been developed for predicting the effects of noise, nonlinear distortion, and linear filtering on speech quality. In HASQI v.1, the extracted features are separately mapped to HI and NH data sets; resulting in a separate indices for each group of listeners. HASQI v.2 [59] is the updated version which has several advantages over the previous version. One of the benefits over the first version is that the auditory periphery model is adjusted to reproduce the effects of HL, so the same quality index can be used for both groups of listeners [59]. The model starts with a representation of the auditory system that incorporates aspects of impaired hearing. One set of features measures the effects of noise and nonlinear distortion on speech quality, whereas second set of features measures the effects of linear filtering. HASQI is the product of the nonlinear and linear subindices.

A study is done in [99] which investigates whether incorporating a model with more physiological details ([98, 101]) compared to HASQI results in greater accuracy in predicting quality for enhanced wideband speech. The aforementioned NSIM measure is utilized in this study to predict speech quality. Parameter optimization is performed to improve the performance of the proposed measure and linear regression is used for mapping purposes.

Modulation Spectrum Area (ModA) [102] is a nonintrusive metric which relies on the fact that reverberation results in smearing the envelope in speech signals. This affects the signal's modulation spectrum. In ModA, the speech signal is first decomposed into 4 acoustic bands with lower cutoff frequencies of 300, 775, 1375, and 3676 Hz. Then by using Hilbert transform, the temporal envelopes in each frequency band are computed. The envelopes are subsequently downsampled and grouped using a one-third-octave filterbank. 13 modulation filters cover the 0.5-10 Hz modulation frequency range. Then, the modulation spectrum area is computed for each acoustic frequency band. The averaged area over 4 acoustic bands is used as the final quality index [102].

A modulation spectral representation for nonintrusive quality and intelligibility assessment is represented in [103]. Using modulation spectral insights, an adaptive measure termed speech to reverberation modulation energy ratio (SRMR) is proposed for reverberant and dereverberated speech. SRMR-HA [60] is a modified and extended version of SRMR that incorporates a HL model to be applicable for HA speech quality assessment. Similar to the HASQI procedure, the distorted signal is first passed through a gammatone filter which is adjusted to account for OHC loss parameter. Then the extracted envelope in each channel is input to an 8-channel modulation filter bank with center frequencies of 4.0 Hz, 6.6 Hz, 10.8 Hz, 17.7 Hz, 29.0 Hz, 47.6 Hz, 78 Hz and 128 Hz. The lower four channels mostly contain speech-related components, while the upper four channels are assumed to contain noise or distortion. The feature set consists of the mean and variance of the modulation filterbank output energies. These features are calculated for all the stimuli and the optimal combination of these features are chosen through multiple linear regression analysis to match the subjective scores [60].

A nonintrusive method is proposed in [81] for the quality assessment of noise-suppressed speech. Mel filter bank energies (FBEs) are used as features because of their sensitivity to noise. Using the Mel bandpass filters enhances the perceptually important frequencies. In this work, 13 linearly-spaced and 27 logarithmically-spaced triangular filters are utilized for signal decomposition. The preliminary results in [81] show that higher order statistical properties (skewness and kurtosis) are not as important and lower order moment (mean and variance). Hence, mean and variance of the FBE are used to obtain and 80-dimensional global feature vector. These features are subsequently mapped to MOS by SVR.

The discussed objective quality metrics can be categorized into intrusive and nonintrusive measures for both telecommunication and Audiology applications (see Table 2.1).

2.6 Summary

In this chapter a review of some of the existing objective quality metrics for NH and HI applications was presented. The perceptually-motivated measures were discussed in more details. Three mapping approaches were introduced as candidates for being used in the final objective measure. These models are mainly used to capture the trend in subjective scores and map the features accordingly. The existing methods for feature extraction and cognitive models form the basic idea behind the nonintrusive quality metrics proposed and validated in the next three chapters.

Table 2.1: Objective speech quality measures for telecommunication and Audiology applications

Telecommunication		**Audiology**	
Intrusive	**Nonintrusive**	**Intrusive**	**Nonintrusive**
PESQ [93] POLQA [94] PEMO-Q [95] ViSQOL [97]	P.563 [15] ModA [102] SRMR [103] FBE [81] NISA [92]	HASQI [100] PEMOQ-HI [96]	SRMR-HA [60]

Chapter 3

Auditory Model Based Metrics

3.1 Introduction

In this chapter an overview of HASQI is given as it is an intrusive quality measure that works for both NH and HI groups. The HL model used in HASQI is discussed. Perceptual linear prediction (PLP) as a technique for speech modeling is reviewed. Then a new methodology for extracting two sets of features from speech signals is introduced. It follows by the details on how feature extraction is performed for each feature set.

3.1.1 HASQI

As discussed in Chapter 2, HASQI is an intrusive objective metric for speech quality assessment [59, 104, 100] which has been investigated for multiple NH and HI applications and demonstrats good predictions of speech quality. Since the metric relies on comparing the two signals (e.g. clean and test signals), part of the model is temporal alignment of the signals which removes large delay differences. Consequently, they are inputs to parallel cochlear models. The cochlear model consists of a gammatone filterbank, envelope computation, compression and a low-pass filter in series [Figure 3.1]. A second temporal alignment is performed to remove any remaining timing delays. The two signals then go through the IHC model and the group delay introduced by the filterbank is compensated in each frequency band. Finally, signal features are extracted from the auditory model outputs to derive the quality index.

In HASQI, the gammatone filter bank simulates the auditory filtering behavior of the basilar membrane (BM). OHC damages cause broadening of the filter bandwidth and is included into the model using the following equation:

$$BW_{increase}(i) = 1 + (\frac{OHC_{Loss}(i)}{50}) + 2 \times (\frac{OHC_{Loss}(i)}{50})^6 \tag{3.1}$$

where $OHC_{Loss}(i)$ is the HL due to the OHC damage in i^{th} gammatone channel, with a maximum of 50 dB. As the bandwidth adjustment was only dependent on the OHC loss, the total hearing loss at each filter centre frequency was apportioned between OHC and IHC loss, with roughly 80% of the total loss attributed to OHCs [104]. In addition, the filter bandwidths increased linearly with increasing intensity above 50 dB Sound Pressure Level (dB SPL) [104].

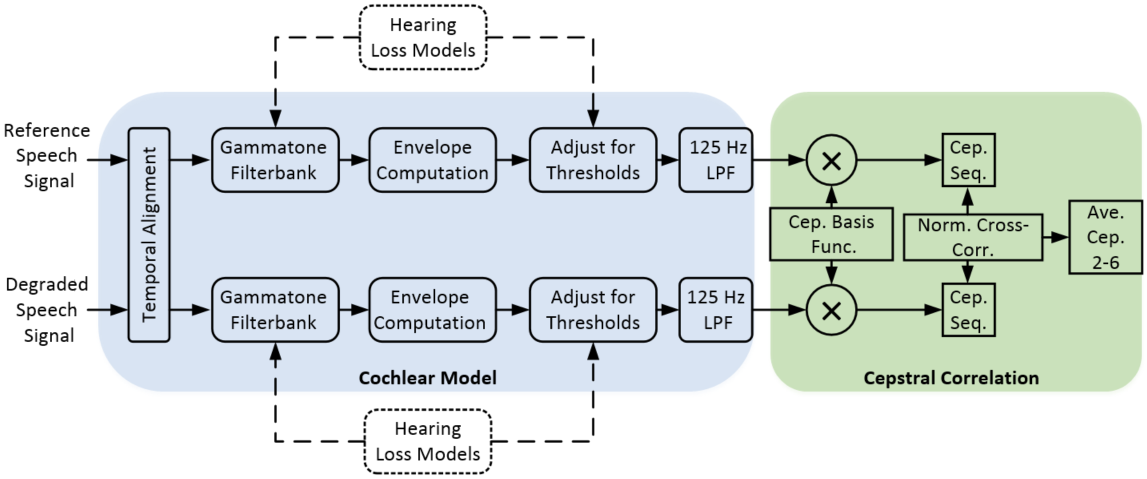

Figure 3.1: Block diagram showing cepstrum correlation procedure for comparing envelopes of clean and test signals.

The compressive behavior of the BM, which is parameterized by the knee point and compression ratio (CR), is also affected by the OHC loss. The envelope is then attenuated in each gammatone frequency channel due to IHC loss. The envelopes are then fitted with a set of six cepstral bases functions and the correlation between the reference and processed envelopes is calculated. The average of the correlations is called cepstrum correlation (CC). A CC value of 1 indicates a processed signal which is perceptually indistinguishable from the clean signal whereas a 0 value indicates a severely distorted signal. Vibration correlation is the second feature extracted for measuring the nonlinear degradation. The normalized cross correlation between BM vibration for the reference and test signals is computed in each frequency band. Each normalized cross-correlation is then weighted by a frequency-dependent weight. The weight is 0 if the segment of clean signal is below auditory threshold and is set to IHC synchronization index for the segments above auditory threshold. The summation of weighted BM vibrations across segments and frequency bands is divided by the sum of the weights to result in the BM vibration.

The nonlinear models based on the cepstral correlation alone and vibration correlation alone is a minimum mean squared error (MMSE) third order regression fit to the combined NH and HI MOS for the noise and distortion stimuli:

$$Q_{Nonlin1} = c^3 \tag{3.2}$$

$$Q_{Nonlin2} = v^3 \tag{3.3}$$

where c is cepstral correlation and v is BM vibration. The third nonlinear model is derived by combining cepstral correlation and BM vibration. This nonlinear model is used in the final HASQI function. It is given by:

$$Q_{Nonlin} = c^2 \times v. \tag{3.4}$$

The linear index accounts for effects on the long term spectrum caused by the HA DSP. By subtracting the normalized input linear signal spectrum from the normalized output signal spectrum, the difference in the spectra is calculated. The difference in the spectral slopes is also computed and the standard deviation of these values is calculated. The linear model is a MMSE linear regression fit to the combined NH and HI subject ratings for the linear filtered stimuli and is given by:

$$Q_{Linear} = 1 - 0.579\sigma_1 - 0.421\sigma_2 \tag{3.5}$$

where σ_1 is the standard deviation of the spectral difference and σ_2 is the standard deviation of the slope difference.

Finally, the combined HA speech quality index is achieved by multiplying the nonlinear index by linear index:

$$Q_{Combined} = Q_{Nonlin} \times Q_{Linear}. \tag{3.6}$$

HASQI has been shown to correlate well with subjective quality ratings by HI listeners across a variety of HA signal processing and environmental conditions [59]. Robustness of the HASQI in predicting subjective quality on databases other than the one on which it was trained was investigated in [62]. In that study, a database of noisy speech suppressed by noise suppression algorithms, along with the subjective scores for NH listeners is used to demonstrate HASQI's prediction performance. The results showed that HASQI works for NH listeners and has performance comparable to other commonly used metrics, however, its performance is poorer compared to the performance reported for the database that HASQI was originally mapped to. While in [59] HASQI was validated with subjective data collected from HI listeners using simulated HA processing, and in non-reverberant environments, The study done by Suelzle [60] validates HASQI on the stimuli recorded through real HAs and in real world environments. Suelzle et al. found that the original regression function used in HASQI to combine the extracted features to the final quality index [100] did not generalize to the new database. The CC values, however, showed a good performance in predicting the quality of speech [60].

3.1.2 Perceptual Linear Prediction

LP is a widely used technique for speech-analysis, modeling, and coding [105]. In LP modeling, the transfer function of the system is typically estimated by an all-pole model as it is relatively simple and straightforward [105]. In PLP, the LP modeling is modified to incorporate the following human perceptual phenomena: 1) the critical-band spectral resolution, 2) the equal-loudness curve, and 3) the non-linear mapping between sound intensity and perceived loudness [106]. It is shown that these modifications enable PLP modeling to be more consistent with human hearing in comparison with conventional LP analysis [106]. It is worthwhile to note here that the order of the PLP model determines the amount of details in the auditory spectrum that is preserved in the PLP model. The higher the order of PLP model, the more speaker-dependent information it captures. A 5-th order all-pole model has been found to be effective in suppressing speaker-dependent details of the auditory spectrum. These properties make the PLP model potentially useful for application to the nonintrusive speech quality assessments. In [107], a nonintrusive metrics for HA applications based on PLP modeling was introduced.

In the PLP-based measure, first the speech signal is framed into 20 msec segments and the Hamming window is applied on speech frames. Then, the windowed speech frames are transformed into the frequency domain and the short-term power spectrum of speech is calculated. In order to apply critical-band analysis, the frequency axis is warped into the Bark frequency scale and the Bark spectrum is convolved with a critical-band filter and sampled at 1-Bark intervals.The sampled power spectrum is then pre-emphasized by the simulated equal-loudness curve. The curve is an approximation to the sensitivity of human hearing at different frequencies. Then, 1/3 power law is applied to simulate the nonlinear mapping between sound intensity and its perceived loudness. The inverse Discrete Fourier transform is applied on the modified spectrum to obtain the corresponding autocorrelation function. The PLP coefficients are determined from Yule-Walker solution of the autocorrelation function. PLP cepstral coefficients are subsequently determined by recursion.

PLP coefficients have previously been shown to perform well in predicting subjective speech quality ratings of narrowband speech coding and noise reduction algorithms [73], and have shown promise for HA speech quality estimation [107]. The promising performance of both PLP and HASQI in predicting perceived speech quality is encouraging in making use of both models in a nonintrusive metric for speech quality assessment.

3.1.3 Perceptual Linear Prediction and Filter Bank Energies Incorporating Hearing Loss Model (PLP-HL and FBE-HL)

The discerning aspects of the proposed approach for deriving the reference-free metrics are two-fold. First, the PLP [106] model-based coefficients were used rather than the LP coefficients and their derivatives as employed in P.563, LCQA-HA, and NISA methods. As discussed above, PLP modeling incorporates normal auditory perceptual phenomena into LP modeling, such as: a) the critical-band spectral resolution, b) the equal-loudness curve, and c) non-linear mapping between sound intensity and perceived loudness [106]. Second, the PLP model was modified to incorporate the effects of sensorineural hearing loss, in a manner similar to HASQI [59]. This step was indeed necessary, as the goal is to predict the HA speech quality as perceived by a HI listener, and the normal PLP model does not take into account the loss of spectral resolution, elevated hearing thresholds, and abnormal growth of loudness that are the unfortunate side-effects of hearing loss. Figure 3.2 depicts the flow chart outlining the computational steps involved in deriving the PLP coefficients while accounting for hearing loss effects (labeled as PLP-HL). The model intakes a properly transformed digital speech sample and the hearing loss information through an Audiogram. The digital Root Mean Square (RMS) value of the speech signal must be scaled to represent the average SPL at the eardrum (a digital RMS value of 1 represented 65 dB SPL). Furthermore, the model operates at a sampling rate of 24 kHz, so the input speech under test was converted to this rate if needed. An antialiasing filter is utilized in this step

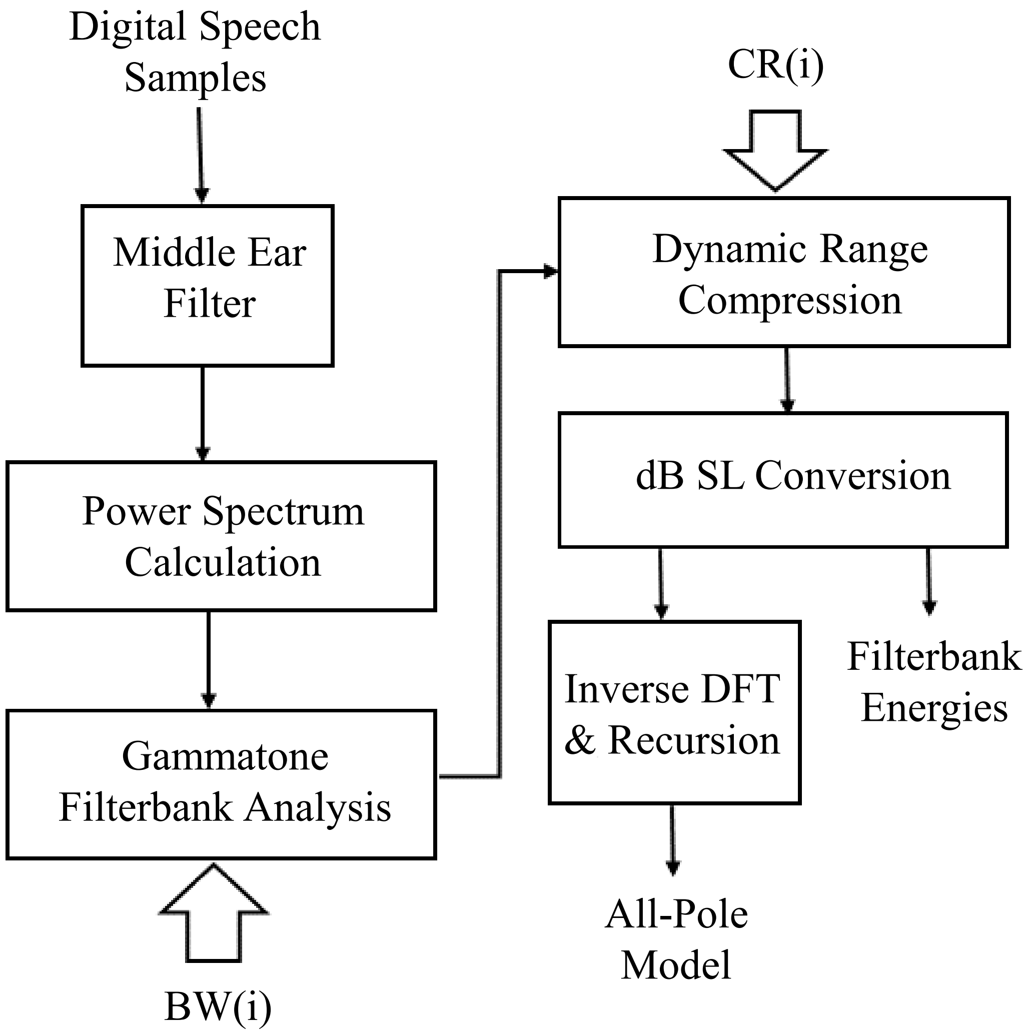

Figure 3.2: Block diagram of PLP-HL and FBE-HL

to avoid aliasing effects on the speech signals. The properly-scaled and transformed speech signal was first put through a middle ear filter, which was modeled as a 2-pole high pass filter at 350 Hz in series with a 1-pole low pass filter at 5000 Hz [104]. The filtered speech was segmented into frames of 20 ms, with each frame windowed by a Hann window and its power spectrum computed. The narrowband power spectra were then multiplied with a matrix of weights representing the gammatone filterbank. The gammatone filterbank was a set of parallel band pass filters, whose implementation was based on Slaney's approach [108]. A total of 32 filters were used in the filterbank matrix, whose centre frequencies were tuned to span the 80 Hz to 8000 Hz frequency range. Each bandpass filter was implemented as an eighth order IIR filter modeling the impulse response of a fourth order gammatone filter [108]. Broadening of the auditory filters as a manifestation of hearing loss was accounted for by adjusting each filter bandwidth relative to that of a normal ear using equation 3.1. The gammatone filter transfer functions for normal hearing are plotted in Figure 3.3 for the filter center frequency range of 80 Hz to 8 kHz. Figure 3.4 depicts the increase in the gammatone filter bank due to an increase in the level of the input signal. A speech signal with an overall level of 100 dB SPL is used to derive the filterbank shown in Figure 3.4. Figure 3.5 shows the BW increase as a result of HL. Maximum HL is specified in this case while the overal level of the speech signal is set to 65 dB SPL.

In the normal PLP model, intensity-loudness mapping for a normal hearing ear was modeled using the cubic-root amplitude compression [106]. However, OHC damage shifts the auditory threshold and reduces the compression ratio in the HI ear. This was modeled by three line segments in which the gain is linear below the lower threshold, compressive with a compression ratio of CR:1 between the lower and upper thresholds, and reverts to linear above the upper threshold [104]. In HASQI, the compression ratio is pre-calculated for each frequency band and is used to weight the envelope of signal after conversion to dB SPL. In PLP-HL, however, the weighted power spectrum of the signal was computed and therefore these computations were applied in frequency domain. The compressed spectrum, after dynamic-range compression, was converted to dB above auditory threshold, i.e. dB Sensation Level (dB SL). In the final stage, autocorrelation coefficients were calculated by taking the inverse Discrete Fourier Transform (DFT) of the modified power spectrum. Finally, Levinson-Durbin recursion was applied to find the all-pole PLP-HL model coefficients. A 6^{th} order PLP model was used in this study, in line with previous studies [e.g. [73]], which have shown that adequate speaker-independent auditory spectral details are preserved with a 6^{th} order PLP model. Cepstral coefficients were subsequently derived from the PLP-HL coefficients through recursion. The PLP-HL feature set included the statistical properties, viz. mean, variance, skewness, and kurtosis of the PLP-HL and cepstral

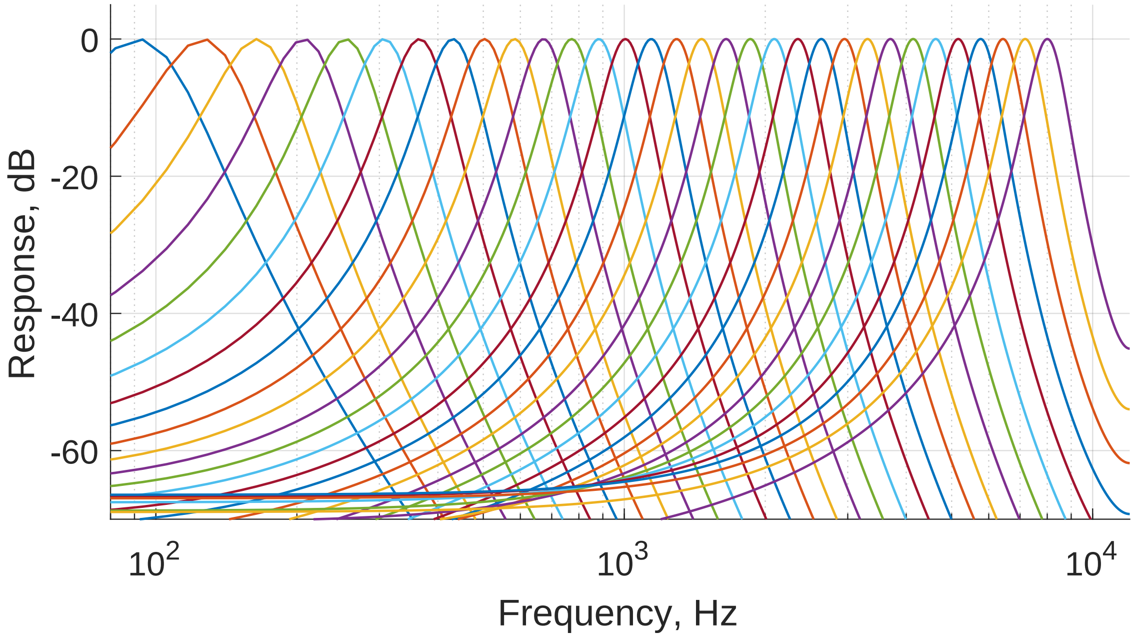

Figure 3.3: Magnitude frequency response of gammatone filters for NH and Nfft = 2048.

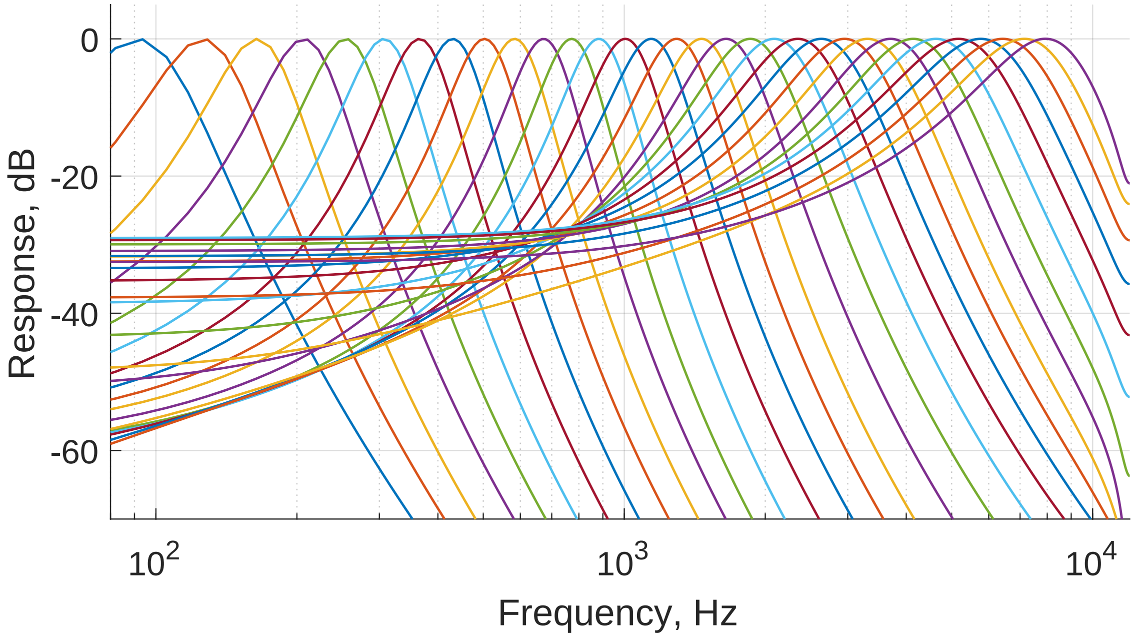

Figure 3.4: Magnitude frequency response of gammatone filters for NH and Nfft = 2048. The BW increase is due to level increase (Overall level = 100 dB SPL).

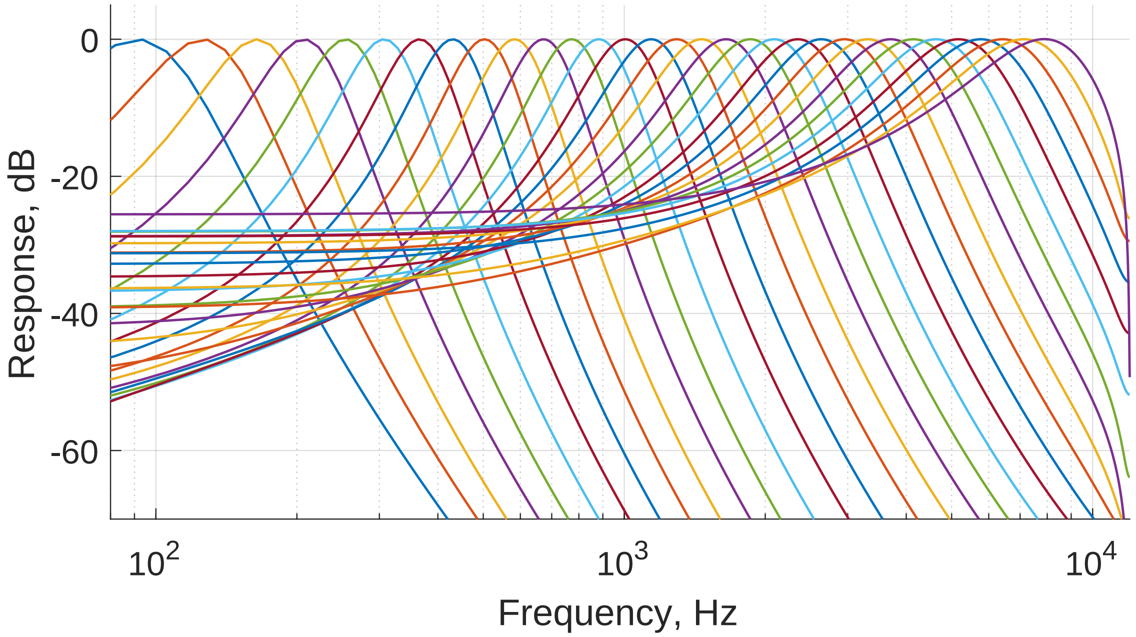

Figure 3.5: Magnitude frequency response of gammatone filters for maximum amount of OHC damage and Nfft = 2048.

coefficients calculated over the entire speech stimulus. The statistical properties along with the absolute value of skewness of the two sets of coefficients resulted in 10 features. By adding the averaged estimated prediction residual error power and spectral entropy, the final number of features in the PLP-HL feature set was 12. In addition to the coefficients extracted in the PLP-HL model, gammatone filterbank energies (labeled as FBE-HL here) were also extracted, as they have been shown to capture cues relevant to quality perception by normal hearing listeners [81]. In FBE-HL, the energies of the weighted spectra in each frequency band after being converted to dB SL were computed. The mean and variance of the frame energies in each of the 32 channels were calculated, which gave rise to a 64 FBE-HL feature set.

As discussed earlier, the extracted features are assimilated through a feature mapping procedure that results in the predicted subjective speech quality score. Three mapping techniques are investigated in this thesis for validation of the proposed feature extraction methods and the results are reported in chapters 4, 5, and 6.

3.2 Summary

In this chapter the HL model used in HASQI is reviewed. PLP model which incorporates perceptual phenomena for NH auditory system is described. Then it is discussed in details how the HL model is incorporated in the PLP technique to account for HI auditory system. Subsequently two sets of features are introduced namely PLP-HL and FBE-HL. The mapping and validation results will be discussed in the following chapters.

Chapter 4

Normal Hearing Speech Quality Assessment

4.1 Introduction

Wideband telephony applications such as cellular and hands free communication devices have rapidly developed and entered the market during the past few years. In such situations, background noise can degrade the quality of speech and as a result noise reduction algorithms are utilized to suppress the noise and deliver a more pleasant speech to the other end. Several noise reduction algorithms have been developed for telecommunication devices and this necessitates benchmarking them. This helps in rank ordering their performance as well as fine-tuning the algorithm to get the maximum benefit.

In order to undertake the quality evaluation, a database containing speech recordings preprocessed by noise reduction algorithms is essential. In this chapter a subjective quality ratings database for wideband noise reduction algorithms is used to validate PLP-HL and FBE-HL. The performance of the proposed metrics is then compared to that of the existing nonintrusive metrics for speech quality evaluation. HASQI is used as an intrusive metric to provide a point of reference in this objective assessment.

4.2 Normal Hearing Database

In order to validate the proposed nonintrusive objective quality prediction measure, a NH database which includes seven different conditions (six wideband NR algorithms and one wideband, unprocessed condition), is used [54].

16 clean speech samples spoken by two male and two female English speakers were used to create the database. The clean speech samples were taken from the telecommunications and signal processing (TSP) speech database [109] with an original sampling rate of 48 kHz which were subsequently down-sampled to 16 kHz. Noisy speech samples were synthesized by mixing clean speech samples with three types of noise (babble, traffic and white) at three SNR levels (0 dB, 5 dB and 15 dB). The noisy speech samples were processed by noise reduction algorithms. Thus, the database contained a total of 1008 enhanced counterparts. The six noise reduction algorithms included three statistical-model-based algorithms (termed *logMMSE*, *logMMSE_SPU*, *WCosh*), one spectral subtraction algorithm (termed *mband*), one Wiener filtering algorithm (termed *Wiener_as*) and one subspace approach algorithm (termed *KLT*). *WB* term refers to the noisy unprocessed stimuli in this database.

Thirty two NH listeners were recruited and each listener rated the quality of all speech stimuli. MUSHRA (MUltiple Stimulus test with Hidden Reference and Anchors) software was used for this experiment [110] and the sound presentation level was adjusted to be at the participant comfortable listening level. Participants were instructed to rate each stimulus using the sliders, by paying particular attention to the clarity, pleasantness, distortion/artifacts, and their overall impression of sound quality.

Pourmand et al. [54] and Wirtzfeld et al. [99] looked at the performance of intrusive quality metrics for this database. However, nonintrusive metrics have not been investigated on this database yet. The purpose of this chapter is to investigate the generalizability of the two proposed feature sets mapped to this database and compare the results with other objective quality metrics.

4.3 Methodology

In order to validate the proposed metrics, they are applied on the NH database along with three other nonintrusive metrics as well as HASQI. LCQA_HA, ModA, PLP-HL, FBE-

HL, and SRMR are the nonintrusive metrics chosen for this study. HASQI is the intrusive metric which has been studied well enough and demonstrated good quality prediction for NH applications.

4.3.1 Feature Extraction

In order to apply PLP-HL algorithm on NH database, the recordings are re-scaled afterwards to reflect their absolute SPL and the input level is specified to be equal to 65 dB SPL. The HL profile is set to 0 dB at the 6 audiometric frequencies (e.g. [250, 500, 1000, 2000, 4000, 6000] Hz). As discussed in Chapter 3, the order of recursion in Levinson-Durbin method is set to 6 and the number of filters in gammatone filterbank is set to 32. Following this procedure, 12-dimensional feature vector is calculated through PLP-HL metric.

In order to compute FBE-HL, the recordings are re-scaled afterwards to reflect their absolute SPL and then fed into PLP-HL algorithm. After the weighted spectrum of each signal frame is calculated, its mean and variance across frames are found. This procedure resulted in 64 feature vector.

In LCQA-HA, each speech recording was segmented into 20 ms non-overlapping frames and an 18^{th} order LPC model was utilized. The individual features and their time derivatives were computed using the LPC model parameters, giving rise to a core set of 10 per-frame features. The global statistical properties of the per-frame features across the entire recording resulted in a 40-dimensional feature vector. In order to calculate the iSNR parameter, the frame-by-frame noise power spectrum was estimated using the minimum statistics algorithm [90]. The frame-specific input and estimated noise power spectra were then grouped into $1/3^{rd}$ octave bands, and the frame $iSNR$ is calculated using equation 2.5, with $I(k)$ representing the $1/3^{rd}$ octave band importance function [90]. Finally, the iSNR values of those frames identified by the VAD algorithm [90] were averaged and the averaged value appended to the above LCQA feature set. By including averaged LPC coefficients skewness and LPC kurtosis parameters into our feature set, a final feature set containing 44 global features was derived from the recordings.

ModA index is calculated by setting the frequency bands to 8 and the envelope frequency to 20 Hz so that it satisfies the $f_{cut} = 10\,Hz$. SRMR is the last nonintrusive metric applied on the recordings and the outputted index is captured for the entire database.

In order to compute HASQI, a reference signal is required. The reference signal associated with each recording is first down-sampled to 16 kHz. Then test signals are time-aligned with the corresponding reference signal, and both signals are given to HASQI v.2 code. The HL profile is set to zero at 6 audiometric frequencies.

4.3.2 Feature Mapping

In order to retain the features that have a high degree of correlation with subjective data, and to minimize redundancy in predictive information carried by different features, two dimensionality reduction approaches are investigated in this work:(a) PCA followed by linear regression, and (b) MARS. Subsequently, the reduced feature set has to be mapped to a single predicted speech quality score for the speech sample under test.

For PCA approach, first, the features were rank-ordered based on the correlation-based index explained in Chapter 2. A subset of the rank-ordered features (ten features) were then transformed using the PCA and combined using a linear regression function.

The MARS method employs forward selection and backward deletion procedures to retain features most important for quality prediction. In this work, the MARS model was built using the ARESLab software [111].

Both PCA and MARS approaches require training data for determining the parameters of the respective mapping function. The data set is randomly divided into 75% and 60% for training the model parameters and the rest were used for prediction using the trained model. This process was repeated 30 times and the averaged correlation coefficient between the predicted and true subjective quality scores in the test-portion of the database was used as the benchmark for performance.

SVR machine is another mapping approach investigated in this work. SVR alone does not take care of dimensionality reduction. Hence, we investigated its performance in three ways: (a) with no dimensionality reduction (whole database was given to SVR machine), (b) PCA-selected feature set, and (c) MARS-selected feature set.

Then cross-validated SVR was applied to the entire database to capture the SVR machine parameters which result in the best performance. By utilizing the $kfold$ parameter in SVR machine, the data is split into k partitions, one partition is used for testing and the remaining $(k-1)$ partitions are used for training. The experiment is repeated with each of the k chunks used for testing. The average of the accuracy of the tests over the k chunks is taken as the performance measure. For this database $kfold$ of 10 resulted in the highest performance. The procedure was repeated for the whole database as well as PCA-selected and MARS-selected features for $linear$, *Gaussian*, and $radial\ basis\ function$ (*rbf*) kernel functions.

It should be highlighted that ModA, SRMR, and HASQI did not use training while the proposed metrics and LCQA-HA require training. Hence, in order to have a fair comparison, it is crucial to test the generalizability of the metrics that require training with varying training and test data. To do so, the database is partitioned in different ways and the performance of the proposed metrics is examined on the untrained data. Four tests are performed to achieve this. *Test 1* includes training a model on a subset of sentences and testing on the remaining sentences. In *SENT12*, sentence 1 to sentence 12 in each condition is selected to be visible by the model and the rest will serve as the unseen data. In *SENT10*, the sentence 7 to sentence 16 in each condition will be visible to the model. *Test 2* includes selecting the training and test sets according to the NR algorithm. In *NR4* test, *logMMSE*, *logMMSE_SPU*, *WCosh*, and *mband* are used for training and the remaining three algorithms are left for testing. In *NR5*, *WCosh*, *mband*, *KLT*, *wiener_as*, and *WB* are used for training and the remaining two algorithms are used for testing. *Test 3* is performed on different SNRs where *SNR05* refers to the test where training set includes SNRs of 0 dB and 5 dB while in *SNR515*, SNRs of 5 dB and 15 dB are set aside for training. *Test 4* is performed on different noise types. In *NOISBT*, babble and traffic noise types form the training set while in *NOISETW*, traffic and white noise types are used for training and babble noise is used for testing. Other combinations are also possible for partitioning the database into training and test sets but they are not explored in this work.

The experimental results are reported in terms of the three criteria commonly used for performance comparison namely: Pearson linear correlation coefficient (for prediction accuracy), standard deviation of the prediction error [55], and Steiger's Z-test [57] (for the statistical significance of the difference among results).

Table 4.1: Correlation coefficient of the predicted data with MOS for the three objective quality measures fitted by PCA and MARS approaches for 75% ratio.

Mapping Approach	**Objective Measure**	ρ		
		Training Set	**Test Set**	**Complete Set**
PCA	LCQA-HA	0.42	0.34	0.40
	FBE-HL	0.40	0.36	0.39
	PLP-HL	0.54	0.53	0.54
MARS	LCQA-HA	0.57	0.51	0.55
	FBE-HL	0.60	0.55	0.59
	PLP-HL	0.60	0.58	0.59

4.4 Results and Discussion

LCQA-HA, FBE-HL, and PLP-HL feature sets were reduced by PCA and then they were linearly fitted to MOS. MARS was also applied on the three measures and the averaged correlations are reported in Table 4.1. It should be noted that PCA and MARS select features based on different criteria (see section 2.3), so although some feature are picked up by both techniques, there are some differences in the selected features. The per-sample correlations are obtained with 75% and 60% of the whole data randomly chosen for training set and the remaining for test set. The results for 60% ratio were either similar or had no statistical significance so only the results for 75% ratio are reported here. MARS approach results in better performance compared to PCA in all cases (see Table 4.1). This could be due to the fact that it uses a more complex way of choosing the best feature set as well as utilizing a better mapping procedure compared to PCA. On average, MARS chooses 10 features from LCQA-HA feature set and the same number of features out of 64 features in FBE-HL measure. For PLP-HL measure, 8 out of 12 features were selected by MARS.

In order to apply SVR, the parameters through cross validated method were captured. It is a good practice to standardize the data since standardization makes predictors insensitive to the scales on which they are measured. It is done by centering and scaling each column of the predictor data by the weighted column mean and standard deviation, respectively. *Linear* , *Gaussian*, and *radial basis* kernel functions were then investigated for this database and the latter two showed somewhat similar results to each other and higher compared to the *linear* kernel function. $kfold$ value of 10 resulted in better results for all cases. The average of the accuracy of the tests over the ten chunks is taken as the performance measure and is reported in Table 4.2. The results show that in all cases feeding the whole feature set to SVR results in a better performance indicating that although the features are correlated and contain redundant information, such redundancy is useful.

Table 4.2: Correlation coefficient of the predicted data with MOS for the three objective quality measures fitted by SVR.

Mapping Approach	Objective Measure	Feature Set	ρ		
			linear	***Gaussian***	***rbf***
SVR	LCQA-HA	Whole	0.49	0.22	0.23
		PCA	0.35	0.42	0.42
		MARS	0.37	0.51	0.51
	FBE-HL	Whole	0.55	0.64	0.63
		PCA	0.48	0.56	0.56
		MARS	0.41	0.66	0.66
	PLP-HL	Whole	0.52	0.65	0.67
		PCA	0.48	0.64	0.63
		MARS	0.49	0.66	0.65

As the final step, SVR and MARS are applied on varying training and tested on the rest of the data. Although the cross-validated SVR suggests that *Gaussian* and *radial basis* functions give better results, both kernel functions resulted in over-fitting for all four tests described earlier. Over-fitting is a common problem in machine learning that occurs when the trained model performance is high (close to 1) but it has a poor correlation (close to 0) on the test data. For example, in *test 1* (*SENT12*) when a model is trained on LCQA-HA features using either *radial basis* or *Gaussian* kernel function , the correlation coefficient of the training set with subjective scores is 0.98. However, when this model is applied on the test portion, the performance is very poor ($\rho = 0.09$). Also, the same procedure on FBE-HL feature set results in correlation coefficient of 0.97 on the training set and 0.08 on the test set for the same test. The reason could be that the results in Table 4.2 are obtained by partitioning the data into 10 chunks (*kfold* = 10). This means the models were trained on nine-tenth of the data and tested on one-tenth of the data which is different from the partitioning for the varying-partitioned tests performed later. This over-fitting does not happen when *linear* kernel function is used, so from now on all SVR mappings utilize *linear* kernel function. The results for the *linear* kernel function in Table 4.2 indicate that for both PLP-HL and FBE-HL measures, SVR and MARS applied on whole feature set perform better. This could be due to the fact that both SVR and MARS are complex machine learning approaches and this enables them to combine the features and extract useful information to find better mapping to MOS. After comparing the results in Table 4.1 and 4.2, no significant advantage is seen for SVR approach over MARS; so both fitting approaches are investigated here.

In order to do correlation analysis, the average score at each condition is used. To obtain the condition-averaged scores, after calculating the objective scores for all speech samples, the scores of all 16 speech samples at a specific condition are averaged. The correlation coefficients of the predicted data for training, test and complete sets as well as condition-averaged correlations for SVR and MARS fitting approaches for *Test 1*, *Test 2*, *Test 3*, and *Test 4* are reported in Table 4.3, 4.4, 4.5, and 4.6, respectively.

It can be concluded from Table 4.3 that except for LCQA-HA, training on 12 sentences and testing on 4 sentences results in a higher performance indicating that both approaches train a better model when they see more data. The same reasoning is valid for the results reported in Table 4.4 where *NR5* which can see 71% of data performs better in all cases compared to *NR4* which trains a model on only 57% of data. Table 4.5 suggests that when the training methods see the high SNR recordings (5 and 15 dB), they perform better in predicting the quality for the low SNR data (0 dB). From Table 4.6, it can be concluded that training a model on *traffic* and *white* noise types results in a better performance. Considering the fact that white noise spectrum is wide whereas traffic and babble noise types have a narrower spectrum, it is logical that a model trained on stimuli corrupted by a wide noise type is generalizable on stimuli with narrower noise types. The difference in performance is more obvious when LCQA-HA feature set is examined. The reason could be that in LCQA-HA features are calculated from LP coefficients which represent a compact form of spectrum and as a result may not carry much information on higher frequency portion of the stimuli when only trained with narrower band noise spectra (traffic and babble).

For a visual comparison, the condition-averaged correlations for the above-mentioned tests are plotted in Figure 4.1. From Figure 4.1, it can concluded that FBE-HL fitted by SVR results in a higher performance to that of MARS-fitted (*SNR05* is an exception). However, performing Z-test on the correlation coefficients shows that not all differences are statistically significant. In fact, the differences seen between SVR and MARS performance for SENT12, NR4, NR5, SNR515, and NOISTW tests on FBE-HL and NOISBT test on PLP-HL are not statistically significant. For PLP-HL measure, however, a model trained by MARS gives better results (except for SNR05). No specific pattern is seen for LCQA-HA measure.

Since partitioning by sentences gives more robust results for all feature sets, the scatter plots for that test are plotted in Figure 4.2. Scatter plot of the combined index in HASQI for condition-averaged analysis is included for comparison. As can be seen from the plots, there is a linear relationship between the objective and the subjective scores.

Mean of PLP-HL coefficients with per-sample correlation coefficient of -0.34 and normalized prediction error and kurtosis of cepstral coefficients with per-sample correlation

Table 4.3: Performance of SVR and MARS models on three objective quality measures for *Test 1*

Mapping Approach	Objective Measure	Training Data	ρ			
			Train	**Test**	**Complete**	**Cond.-averaged**
SVR	LCQA-HA	*SENT12*	0.60	0.22	0.17	0.66
		SENT10	0.60	0.34	0.49	0.76
	FBE-HL	*SENT12*	0.62	0.37	0.41	0.80
		SENT10	0.66	0.14	0.28	0.80
	PLP-HL	*SENT12*	0.54	0.56	0.53	0.72
		SENT10	0.54	0.47	0.50	0.66
MARS	LCQA-HA	*SENT12*	0.59	0.17	0.23	0.66
		SENT10	0.59	0.45	0.53	0.76
	FBE-HL	*SENT12*	0.63	0.43	0.51	0.77
		SENT10	0.63	0.42	0.46	0.71
	PLP-HL	*SENT12*	0.60	0.59	0.59	0.78
		SENT10	0.59	0.52	0.55	0.73

Table 4.4: Performance of SVR and MARS models on three objective quality measures for *Test 2*

Mapping Approach	Objective Measure	Training Data	ρ			
			Train	**Test**	**Complete**	**Cond.-averaged**
SVR	LCQA-HA	*NR4*	0.60	0.22	0.36	0.48
		NR5	0.62	0.28	0.43	0.60
	FBE-HL	*NR4*	0.62	0.45	0.52	0.71
		NR5	0.63	0.52	0.58	0.75
	PLP-HL	*NR4*	0.55	0.37	0.44	0.55
		NR5	0.57	0.41	0.52	0.69
MARS	LCQA-HA	*NR4*	0.67	0.20	0.40	0.51
		NR5	0.58	0.32	0.49	0.66
	FBE-HL	*NR4*	0.59	0.51	0.56	0.68
		NR5	0.64	0.50	0.58	0.73
	PLP-HL	*NR4*	0.62	0.44	0.53	0.68
		NR5	0.65	0.48	0.59	0.79

Table 4.5: Performance of SVR and MARS models on three objective quality measures for *Test 3*

Mapping Approach	Objective Measure	Training Data	ρ			
			Train	**Test**	**Complete**	**Cond.-averaged**
SVR	LCQA-HA	*SNR05*	0.59	0.03	0.36	0.46
		SNR515	0.52	0.30	0.50	0.64
	FBE-HL	*SNR05*	0.57	0.28	0.45	0.54
		SNR515	0.58	0.1	0.51	0.75
	PLP-HL	*SNR05*	0.53	0.13	0.35	0.44
		SNR515	0.53	0.18	0.53	0.65
MARS	LCQA-HA	*SNR05*	0.58	0.00	0.32	0.39
		SNR515	0.55	0.25	0.53	0.71
	FBE-HL	*SNR05*	0.56	0.31	0.52	0.63
		SNR515	0.58	0.26	0.57	0.71
	PLP-HL	*SNR05*	0.60	0.13	0.31	0.33
		SNR515	0.58	0.16	0.55	0.71

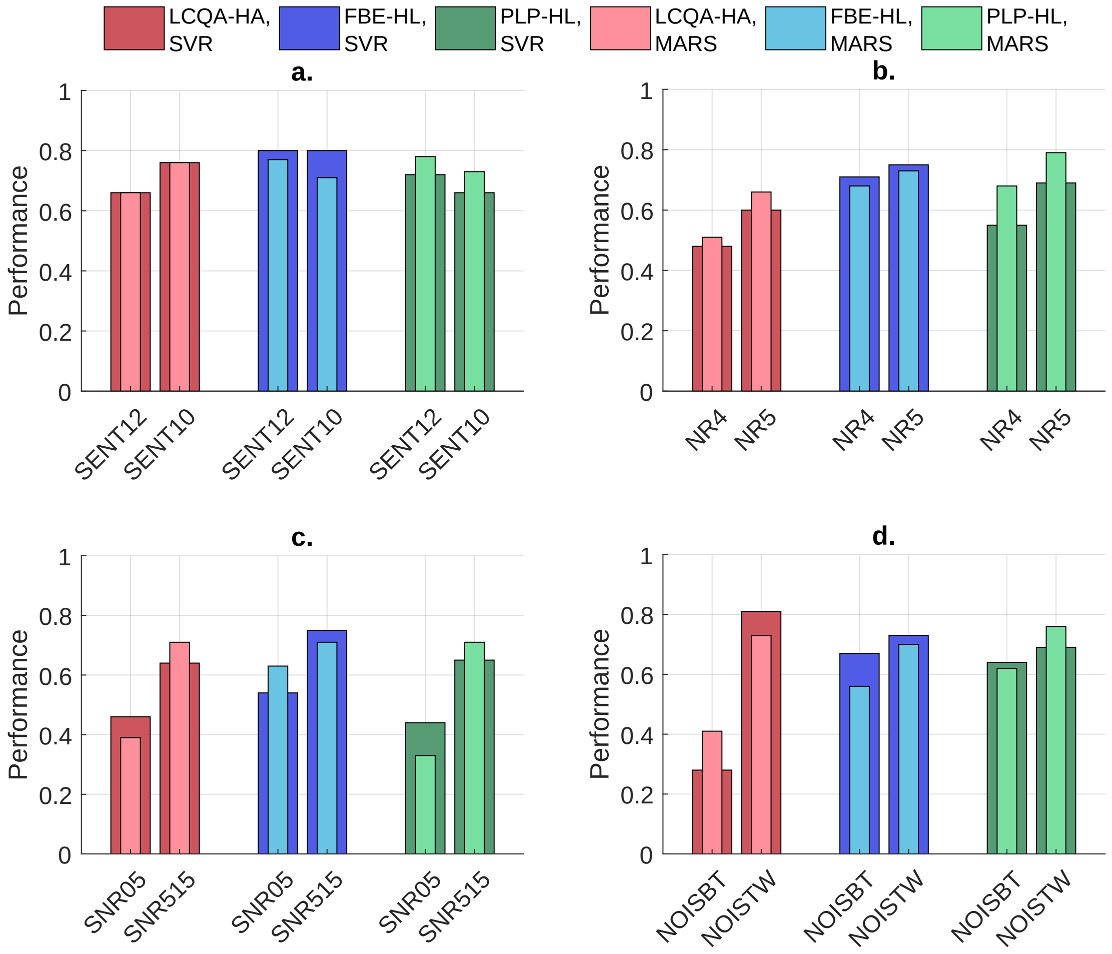

Figure 4.1: Correlation coefficients (performance) of condition-averaged analysis for LCQA-HA, FBE-HL, and PLP-HL with different partitioning of data into training and test sets: a. *Test 1* b. *Test 2* c. *Test 3* d. *Test 4*

Table 4.6: Performance of SVR and MARS models on three objective quality measures for *Test 4*

Mapping Approach	**Objective Measure**	**Training Data**	*ρ*			
			Train	**Test**	**Complete**	**Cond.-averaged**
SVR	LCQA-HA	*NOISBT*	0.62	0.21	0.23	0.28
		NOISTW	0.59	0.44	0.54	0.81
	FBE-HL	*NOISBT*	0.66	0.38	0.52	0.67
		NOISTW	0.62	0.39	0.55	0.73
	PLP-HL	*NOISBT*	0.54	0.50	0.51	0.64
		NOISTW	0.54	0.52	0.54	0.69
MARS	LCQA-HA	*NOISBT*	0.60	0.33	0.31	0.41
		NOISTW	0.56	0.40	0.52	0.73
	FBE-HL	*NOISBT*	0.60	0.29	0.46	0.56
		NOISTW	0.60	0.52	0.58	0.70
	PLP-HL	*NOISBT*	0.64	0.50	0.53	0.62
		NOISTW	0.63	0.46	0.58	0.76

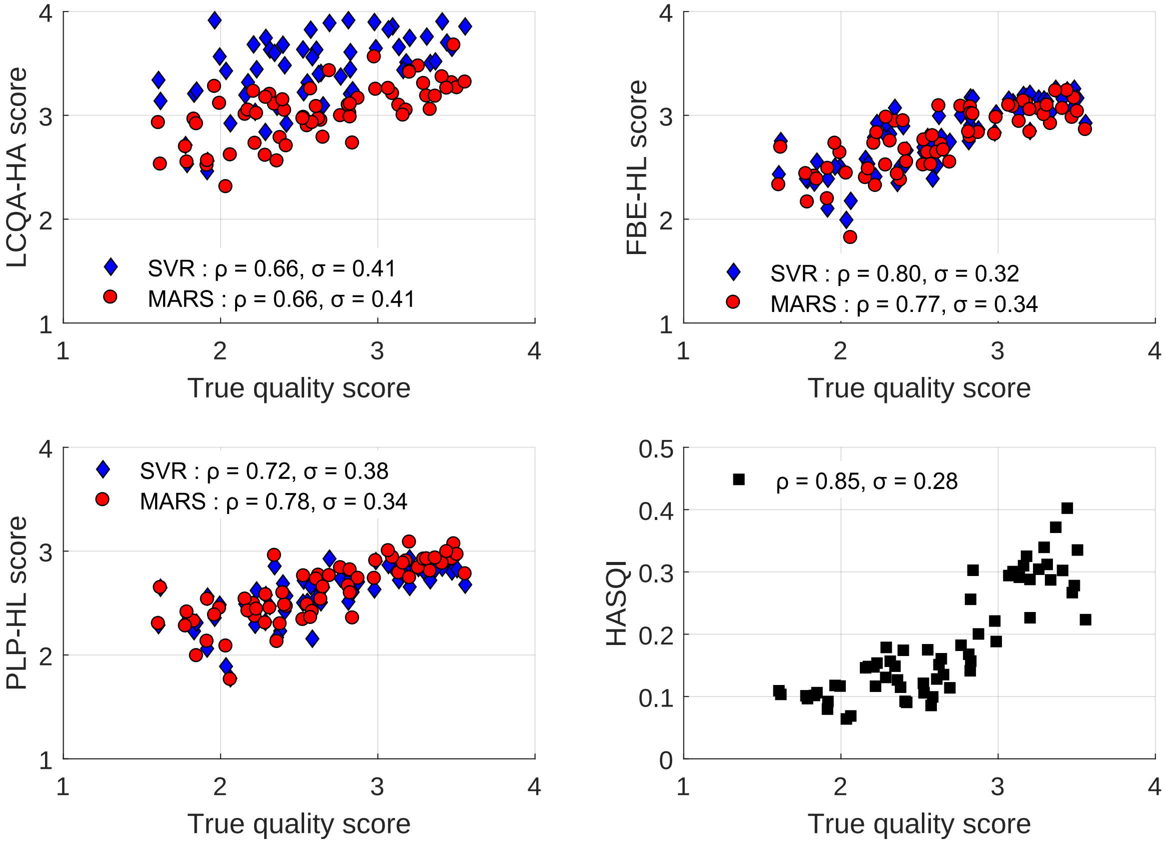

Figure 4.2: Relationship between true and predicted quality scores for LCQA-HA, FBE-HL, and PLP-HL measures fitted by SVR and MARS and HASQI in *Test 1*: condition-averaged analysis

coefficient of -0.24 score higher among all 12 features extracted in PLP-HL method. The condition-averaged correlations for these features are -0.44, -0.25, and -0.29 respectively. After being fitted to MOS, the per-sample correlations vary from 0.31 to 0.59 for per-sample analysis and from 0.33 to 0.79 for condition-averaged analysis.

Mean(spectral flatness) and skewness(signal variance) with per-sample correlations of -0.24 and -0.20 are the highest correlations in LCQA-HA metric. The condition-averaged correlations for these two features are -0.28, -0.47. After being fitted to MOS, the per-sample correlations vary from 0.17 to 0.54 for per-sample analysis and from 0.28 to 0.81 for condition-averaged analysis.

Table 4.7: Correlation coefficient and standard deviation of error for different objective quality metrics for per-sample and condition-averaged analysis

Objective Measure	**Complete Database**		**Condition-averaged**	
	ρ	σ	ρ	σ
ModA	0.1	0.63	0.14	0.54
SRMR	0.1	0.63	0.28	0.52
HASQI-CC	0.60	0.51	0.83	0.30
HASQI-nonlinear	0.65	0.48	0.85	0.28
HASQI	0.67	0.47	0.85	0.28

Averaged energies of the third and fourth frequency bands in FBE-HL metric show correlations of 0.32 and 0.34 for per-sample analysis and 0.36 and 0.52 for condition-averaged analysis. After being fitted to MOS, the per-sample correlations vary from 0.28 to 0.58 for per-sample analysis and from 0.54 to 0.80 for condition-averaged analysis.

Finally, performance of ModA, SRMR, and HASQI indices is reported in Table 4.7. From Table 4.7, SRMR results in 0.1 correlation for per-sample analysis and 0.28 for condition-averaged analysis. ModA measure has per-sample correlation of 0.1 and it is increased to 0.14 after condition-averaging. In HASQI, cepstrum correlation has per-sample correlation of 0.6. After being combined with vibration correlation and mapped to the nonlinear index, the correlation increases to 0.65. The combined index has per-sample correlation of 0.67. After condition-averaging these correlations are increased to 0.83, 0.85 and 0.85 respectively. It can be seen from the results in Table 4.4 that ModA and SRMR both fail at predicting quality of speech for this database. HASQI results in relatively low correlation for the complete database; however, when sentence variability is eliminated, it shows high correlations with MOS. Among all three feature sets, FBE-HL trained by SVR in *test 1* has performance comparable to HASQI. This indicates that it is crucial for the learning techniques to see all the conditions for training a model. It should be noted that no optimization was applied on ModA, SRMR and HASQI-CC indices whereas both nonlinear and combined HASQI indices are optimized by regression analysis.

4.5 Summary

This chapter was on validating the proposed objective speech quality metrics on a NH database. LCQA-HA was also investigated on the same database. HASQI was applied as an intrusive metric. The proposed metrics along with LCQA-HA were mapped to MOS and their performance was analyzed. FBE-HL feature set shows better performance compared to the other two feature sets. Its performance is comparable to that of HASQI when the model sees all the conditions. The results show that FBE-HL and PLP-HL are applicable for future testing of NR algorithms for wideband telephony.

Chapter 5

Hearing Impaired Speech Quality Assessment

5.1 Introduction

According to World Health Organization, over 360 million people have disabling HL [18]. HAs form the most common treatment modality for HI individuals. However, satisfaction with HAs is an important factor in HA adoption. A new survey reports that the first driver of HA satisfaction centers on sound quality [19]. As such, assessment and monitoring of HA speech quality is of significant importance to HA designers, Audiologists, and researchers. HA speech quality is not only affected by environmental influences like background noise and reverberation, but also by distortions caused by the DSP algorithms within these devices. It is therefore important to quantify the impact of HAs and their DSP algorithms on perceived speech quality under a variety of environmental operating conditions.

In this chapter, two data sets containing audio recordings processed by different HAs are utilized to validate PLP-HL and FBE-HL. The performance of the proposed metrics is then compared to that of the existing nonintrusive metrics for speech quality evaluation. HASQI serves as a point of reference for the first database and used to estimate MOS for the second database where subjective scores are not available.

5.2 Hearing Impaired Database

The first HA database used in this study has been previously used in benchmarking HASQI and SRMR-HA metrics [112] and is briefly described here. A group of 18 HI listeners were recruited to provide the speech quality ratings. The HL profile of all the participants was similar between the left and right ears, and the HL severity ranged from mild to severe. Figure 5.1 displays the averaged left and right ear audiograms, as well as the maximum and minimum thresholds across the seven audiometric frequencies. It is evident from this figure that on average the participants had symmetric HL and the severity ranged from mild to severe. In order to collect the subjective quality ratings, a speech database was created which consisted of recordings from the experimental HAs under a variety of environmental conditions. The recordings were obtained using two different bilateral wireless HA models, viz. Oticon Agil and Siemens Motion, which were programmed to fit to the specific HL of each study participant and placed on a head and torso simulator (HATS). The HATS was then positioned at the centre of a loudspeaker array, either in a hemi-anechoic chamber ($RT_{60} = 40\,ms$) or in a reverberant chamber ($RT_{60} = 890\,ms$). In each of these environments, IEEE speech sentences spoken by a male and a female talker were played back from 0° degree azimuth. Three different noise types viz. multi-talker babble, traffic noise and speech-shaped noise (SSN) at overall SNRs of 0 dB and 5 dB were played back from speakers at 90°, 180°, and 270°, to simulate different noisy conditions. The SNRs of 0 dB and 5 dB were chosen as they represent realistic sound environments encountered by HA users [113]. Male speech samples were recorded with the first two types of noise, while female speech samples were recorded with all three types of noise. Using this procedure, a total of 160 conditions (80 each in low reverberation and high reverberation environments) were simulated for HA quality assessment [112]. Under each of these conditions, the stereo HA recordings were sampled at 16 kHz and stored separately for each HI participant. The recorded stimuli were later presented through insert earphones to the respective HI participants, and their ratings of HA speech quality in different conditions were obtained using the MUSHRA procedure. The interested reader is referred to [112] for further details and statistical analysis of the HA speech quality ratings database.

The second database [112] used for this study consisted of speech combined with various types of noise. The HAs used to create this database were loaned through the Hearing Instrument Review Committee (HIRC) program and this name is used to refer to the second database in this thesis. While binaural database comprised of bilateral HA recordings, HIRC database consisted of unilateral HA recordings collected in a desktop HA test box. This configuration was selected for its clinical relevance, as test boxes are typically used

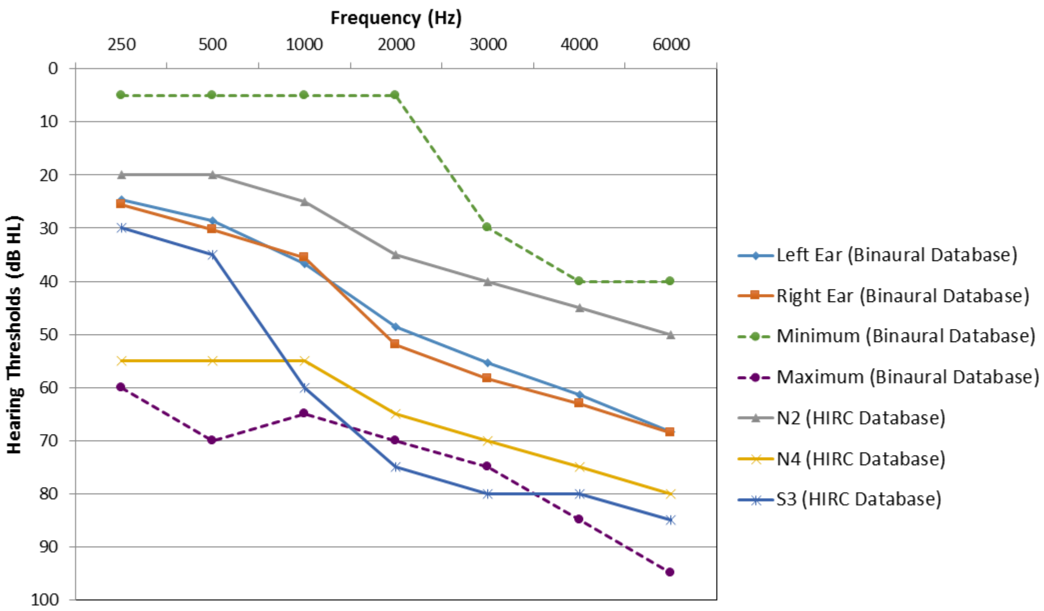

Figure 5.1: Audiograms used in programming the HAs for binaural and HIRC databases. The minimum and maximum thresholds across both ears is also reported for binaural database.

for electroacoustic verification of HA functionality in an Audiology clinic. Seven different HAs, viz. Siemens Motion, Oticon Agil, Starkey S Series iQ, Phonak, Unitron Passport, Widex M440-9, and Sonic Innovations Velocity (randomly assigned HA1-HA7 labels), were used for the creation of this database. The HAs were programmed to fit each of three standard audiograms; a moderately sloping mild loss (labeled as N2), a steeply sloping moderate/severe loss (labeled as S3) and a moderately slopping moderate/severe loss (labeled as N4) as defined in Bisgaard et al. [114] (see Figure 5.1). Each HA was in turn connected to an IEC 126 2-cc coupler and placed within an Interacoustics dedicated test chamber-TBS25 M/P. The International Speech Test Signal (ISTS) [115] at 65 dB SPL was chosen as the speech input to the HAs, as it is widely used in clinical HA test boxes and has been employed in speech quality assessment studies with HI listeners [116]. Three types of noise, viz. SSN, multi-talker babble, and traffic noise were added to the ISTS signal separately at 0 dB and 5 dB SNR. The noisy signal was played back through an internal speaker and the HA response through the coupler was recorded and stored. To evaluate the functioning of the NR algorithms in the seven HAs, recordings were obtained when the NR was enabled and when the NR was disabled. This resulted in a total of forty-two recordings per HA. Unlike the binaural database, no subjective ratings were collected for the HIRC database. Rather, an objective, full-reference metric (HASQI) served as a surrogate for the subjective scores in training the feature mapping function for the reference-free metrics.

5.3 Methodology

ModA, SRMR-HA, LCQA-HA and HASQI are applied on binaural and HIRC databases along with the two proposed metrics. LCQA-HA, ModA and SRMR-HA are the nonintrusive metrics chosen for this study. HASQI is the intrusive metric which has been studied well enough and demonstrated good quality prediction for HA speech quality evaluation.

In order to validate the proposed metrics and examine their robustness, different partitioning is applied on data and the trained models are subsequently tested on the unseen data. In the end, the models trained on binaural and HIRC databases are applied on the other database and results are reported.

5.3.1 Feature Extraction

In order to apply PLP-HL algorithm on binaural database, the recordings are first high pass filtered and re-scaled afterwards to reflect their absolute SPL. The HL profile is set to that of measured for left and right ears separately. The order of recursion in Levinson-Durbin method is set to 6 and the number of filters in gammatone filterbank is set to 32. Following this procedure, 12-dimensional feature vector is calculated through PLP-HL metric.

In order to extract LCQA-HA features, the same procedure as outlined in chapter 4 is followed. ModA index is calculated by setting the frequency bands to 8 and the envelope frequency to 20 Hz so that it satisfies the $f_{cut} = 10Hz$. In order to compute FBE-HL, the recordings are re-scaled afterwards to reflect their absolute SPL. The weigheted spectrum is compressed and then the compressed spectrum is converted to dB SL. Finally, mean and variance of the energy in each of 32 frequency bands is calculated to give rise to 64 features. SRMR-HA is the last nonintrusive metric applied on the recordings and the outputted index is captured for the entire database.

HASQI requires a reference signal to predict the quality of signal under test. Before any comparison takes place, the reference signal should be re-scaled to reflect its absolute SPL, and then frequency shaped using real ear aided gains (REAGs). These gains are calculated for each individuals left and right ear separately by the Desired Sensation Level (DSL) 5.0

algorithm [117]. Then, the signal under test is high pass filtered and normalized and subsequently time-aligned with reference signal. These two signals along with the corresponding HL profile and the level of the input signal are inputted to HASQI version 2 algorithm.

All of the above mentioned measurements are applied for binaural and HIRC databases. However, for binaural database the procedure is repeated for left and right recordings separately and then averaged for each individual. Finally, all features are averaged across individuals.

5.3.2 Feature Mapping

In order to retain the features that have a high degree of correlation with subjective data, and to minimize redundancy in predictive information carried by different features, two dimensionality reduction approaches are investigated in this work:(a) PCA followed by linear regression, and (b) MARS. Subsequently, the reduced feature set has to be mapped to a single predicted speech quality score for the speech sample under test.

For the PCA approach, first, the features were rank-ordered based on the correlation-based index explained in Chapter 2. A subset of the rank-ordered features were then transformed using the PCA and combined using a linear regression function.

The same procedure outlined in Chapter 4 is followed here for obtaining the averaged performance of PCA, MARS and SVR techniques on the two databases. For this database $kfold$ of 10 resulted in the highest performance. The procedure was repeated for the whole feature sets as well as PCA-selected and MARS-selected features for *linear*, *Gaussian*, and *radial basis* kernel functions.

It should be highlighted that ModA, SRMR-HA, and HASQI did not use training while the proposed metrics and LCQA-HA require training. Hence, in order to have a fair comparison, it is crucial to test the generalizability of the metrics that require training with varying training and test data. To do so, the database is partitioned in different ways and the performance of the proposed metrics is examined on the untrained data. Six tests are performed to achieve this. *Test 1* includes training a model on stimuli with SNR of 0 dB in binaural database and testing it on the stimuli with SNR of 5 dB from the same database (*SNR0*). *Test 1* also includes training on SNR 5 dB stimuli and testing on SNR 0 dB recordings (*SNR5*). *Test 2* includes selecting the training and test sets from binaural database according to the type of noise in the stimuli. It is done in two ways, first a model is trained on stimuli containing babble and traffic noise types and then testing on SSN stimuli (*NOISBT*). Second, a model is trained on traffic and SSN stimuli and tested on the recordings with babble noise (*NOISTS*). In *Test 3*, HIRC database is partitioned into recordings with SNRs of 0 dB and 5 dB and one is used as training set while the other one serves as the test set and vice versa (*SNR0* and *SNR5*). *Test 4* examines robustness of the trained models by partitioning the HIRC database based on the HA under test. In *HAOPSS*, the recordings made with Oticon Agile, Phonak Ambra, Siemens Motion, and Sonic Innovations HAs are used for training a model and the model is tested on the recordings made with Starkey S series, Resound Verso, and Widex M440 HAs. In *HASSRW*, a model is trained on the recordings made with Sonic Innovations, Starket S series, Resound Verso, and Widex M440 HAs and tested on Oticon Agile, Phonak Ambra, and Siemens Motion recordings. *Test 5* includes partitioning the HIRC database based on HL profile that HAs were fitted to. Hence, *N4S3* refers to the test where the recordings for N4 and S3 standard audiograms are used for training while N2 recordings are used for testing and *S3N2* refers to the test where S3 and N2 audiograms are used for training and the model was tested on N4 audiogram. Finally, in *Test 6* a model is trained on binaural database and tested on HIRC database (*Binaural*). In *HIRC* test, a model is trained on HIRC databse and tested on binaural database.

The experimental results are reported in terms of the three criteria commonly used for performance comparison namely: Pearson linear correlation coefficient (for prediction accuracy), standard deviation of the prediction error [55], and Steiger's Z-test [57] (for the statistical significance of the difference among results).

5.4 Results and Discussion

5.4.1 Binaural Database

In the same manner as chapter 4, LCQA-HA, FBE-HL, and PLP-HL measures were reduced by PCA and linearly mapped to MOS. MARS was also applied as the second approach and the averaged correlations are reported in Table 5.1. This procedure was done on 75% and 60% of the whole data randomly chosen for training set and the remaining for test set. The results for 60% ratio were either similar or had no statistical significance so only the results for 75% ratio are reported here.

In order to apply SVR, the parameters through cross validated method were captured. The data is standardized and then *linear*, *Gaussian*, and *radial basis* kernel functions were investigated. Utilizing *linear* kernel function resulted in either better or similar (no statistically significant) results to that of the other two functions (See Table 5.2). $Kfold$ value of 10 gave better results for all cases. The average of the accuracy of the tests over the ten chunks is taken as the performance measure and reported in Table 5.2. The results show that in all cases feeding the whole feature set to SVR results in a better performance. When looking at Tables 5.1 and 5.2, FBE-HL outperforms both LCQA-HA and PLP-HL measures when linearly fitted by PCA approach; however, it does not show any statistically significant improvement over the other two measures when mapped by either MARS or SVR.

As the final step, SVR models are trained on varying training and tested on the rest of the data. *Linear* kernel function is utilized as the other two functions resulted in over-fitting for all the tests. After comparing the results in Table 5.1 and 5.2, no significant advantage is seen for SVR approach over MARS, so both fitting approaches are investigated here.

The correlation coefficients of the predicted data for training, test and complete sets for SVR and MARS fitting approaches are reported in Tables 5.3, 5.4.

Tables 5.3 and 5.4 suggest that no one measure outperforms the other two for the analyses performed on the complete data set. There is an exception for this observation in *Test 1* for PLP-HL measure trained on *SNR5* data where the correlation for complete data set equals 0.81, which is statistically significant compared to the other two objective measure fitted on the same training data.

Figure 5.2 shows the relationship between true quality score (MOS) and predicted quality score for the proposed metrics as well as LCQA-HA and HASQI for *Test 2*. A better

Table 5.1: Correlation coefficient of the predicted data with MOS for the three objective quality measures fitted by PCA and MARS approaches: binaural database

Mapping Approach	**Objective Measure**	ρ		
		Training Set	**Test Set**	**Complete Set**
PCA	LCQA-HA	0.93	0.92	0.93
	FBE-HL	0.95	0.94	0.95
	PLP-HL	0.91	0.90	0.91
MARS	LCQA-HA	0.96	0.92	0.95
	FBE-HL	0.96	0.93	0.96
	PLP-HL	0.95	0.92	0.94

metric would show less scatter around the 45° line. This figure shows that LCQA-HA and FBE-HL perform similarly when mapped by either SVR or MARS (*Test 2*). For PLP-HL measure, MARS fitted data results in a statistically significant performance compared to SVR ($Z - score = -4.6$). Performance of combined index in HASQI is poorer compared to the nonintrusive measures; however, it should be noted that HASQI was not trained on this database. The quality scores predicted by HASQI are on the poorer side of quality because the regression function was not trained on stimuli containing reverberation. That is why HASQI predicts poor speech quality for this database.

In Figure 5.4 (a. and b.), correlation coefficients are used to compare the performance of the three measures fitted by MARS and SVR and performed on *Test 1* and *Test 2*. There is no statistically significant difference between SVR and MARS applied on PLP-HL and FBE-HL for these tests.

Table 5.2: Correlation coefficient of the predicted data with MOS for the three objective quality measures fitted by SVR: binaural database

Mapping Approach	**Objective Measure**	**Feature Set**	ρ		
			linear	***Gaussian***	***rbf***
SVR	LCQA-HA	Whole	0.93	0.86	0.87
		PCA	0.91	0.85	0.89
		MARS	0.73	0.88	0.84
	FBE-HL	Whole	0.94	0.90	0.88
		PCA	0.91	0.92	0.93
		MARS	0.93	0.91	0.93
	PLP-HL	Whole	0.90	0.89	0.89
		PCA	0.88	0.92	0.91
		MARS	0.90	0.89	0.90

Standard deviation of cepstral coefficients and averaged cepstral coefficients with correlation of 0.85 and -0.80, respectively; are the two most significant features in PLP-HL objective measure. After being fitted to MOS, the correlations vary from 0.81 to 0.94 for the performed tests.

Averaged iSNR with correlation coefficient of 0.89 and kurtosis of spectral flatness with correlation coefficient of 0.63 score higher among all features extracted in LCQA-HA measure. After being fitted to MOS, the correlations vary from 0.90 to 0.94 in *Test 1* and *Test 2* analyses.

Standard deviation of energies in 9th, 21st, and 22nd frequency bands in FBE-HL metric have correlations of 0.80, 0.81, and 0.79. After being fitted to MOS, the per-sample correlations vary from 0.86 to 0.94 in the performed tests.

Table 5.3: Performance of SVR and MARS models on three objective quality measures: *Test 1*

Mapping Approach	**Objective Measure**	**Training Data**	ρ		
			Train	**Test**	**Complete**
SVR	LCQA-HA	*SNR0*	0.96	0.88	0.92
		SNR5	0.95	0.88	0.91
	FBE-HL	*SNR0*	0.95	0.90	0.94
		SNR5	0.93	0.90	0.93
	PLP-HL	*SNR0*	0.90	0.86	0.89
		SNR5	0.91	0.83	0.86
MARS	LCQA-HA	*SNR0*	0.97	0.88	0.92
		SNR5	0.95	0.86	0.92
	FBE-HL	*SNR0*	0.98	0.81	0.89
		SNR5	0.93	0.89	0.93
	PLP-HL	*SNR0*	0.97	0.85	0.90
		SNR5	0.94	0.75	0.81

Table 5.4: Performance of SVR and MARS models on three objective quality measures: *Test 2*

Mapping Approach	**Objective Measure**	**Training Data**	ρ		
			Train	**Test**	**Complete**
SVR	LCQA-HA	*NOISBT*	0.95	0.94	0.94
		NOISTS	0.96	0.93	0.90
	FBE-HL	*NOISBT*	0.95	0.95	0.94
		NOISTS	0.96	0.91	0.94
	PLP-HL	*NOISBT*	0.90	0.94	0.90
		NOISTS	0.92	0.91	0.90
MARS	LCQA-HA	*NOISBT*	0.95	0.93	0.94
		NOISTS	0.97	0.81	0.91
	FBE-HL	*NOISBT*	0.96	0.96	0.94
		NOISTS	0.96	0.92	0.90
	PLP-HL	*NOISBT*	0.95	0.93	0.94
		NOISTS	0.96	0.87	0.91

5.4.2 HIRC Database

The same procedure as binaural database is applied on HIRC database. Results for PCA and MARS approaches are reported in table 5.5. While MARS outperforms PCA for LCQA-HA and PLP-HL, FBE-HL is more robust in both cases.

Results for SVR-fitted measures for *linear*, *Gaussian*, and *radial basis* kernel functions are reported in Table 5.6. Again, *Gaussian*, and *radial basis* kernel functions over-fit the data for the partitioning tests. For example when *Test 5* is performed for FBE-HL, an SVR mapping that utilizes *radial basis* function gives a correlation coefficient of 0.96 for

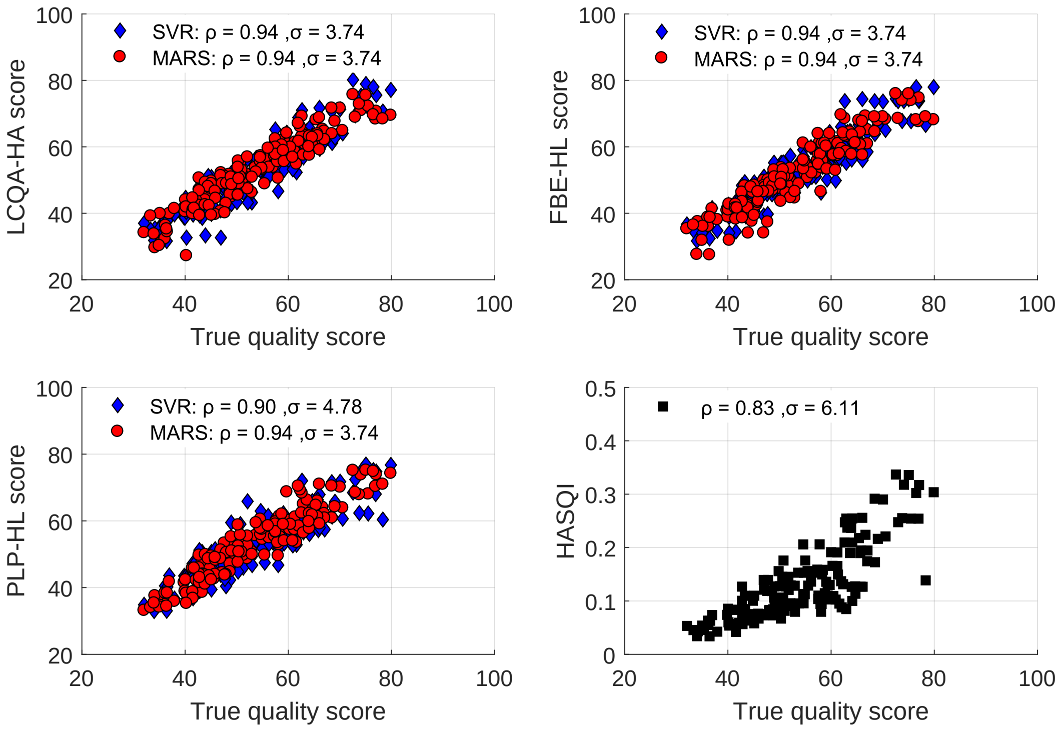

Figure 5.2: Relationship between true and predicted quality scores for LCQA-HA, FBE-HL, and PLP-HL measures fitted by SVR and MARS and HASQI in *Test 2*: Binaural database

Table 5.5: Correlation coefficient of the predicted data with MOS for the three objective quality measures fitted by PCA and MARS approaches: HIRC database

Mapping Approach	**Objective Measure**	ρ		
		Training Set	**Test Set**	**Complete Set**
PCA	LCQA-HA	0.55	0.42	0.52
	FBE-HL	0.91	0.87	0.90
	PLP-HL	0.76	0.74	0.76
MARS	LCQA-HA	0.77	0.55	0.70
	FBE-HL	0.93	0.89	0.92
	PLP-HL	0.82	0.77	0.81

training set (N4S3) and 0.00 for test set. As another example, PLP-HL is investigated for the same test and different training set (S3N2), and the same kernel function results in correlation of 0.93 for the training set and 0.00 for test set. Hence, *linear* function is used for generalizability investigations.

Table 5.6: Correlation coefficient of the predicted data with MOS for the three objective quality measures fitted by SVR: HIRC database

Mapping Approach	**Objective Measure**	**Feature Set**	ρ		
			linear	***Gaussian***	***rbf***
SVR	LCQA-HA	Whole	0.59	0.27	0.28
		PCA	0.38	0.31	0.28
		MARS	0.39	0.48	0.50
	FBE-HL	Whole	0.93	0.60	0.63
		PCA	0.86	0.86	0.84
		MARS	0.83	0.91	0.91
	PLP-HL	Whole	0.76	0.72	0.69
		PCA	0.77	0.75	0.75
		MARS	0.75	0.76	0.75

Test 3 is performed by partitioning the data based on SNR. Results for *Test 3* are reported in Table 5.7. For this test, 50% of data is used for training a model and the rest is used for testing. In almost all cases in this test, the correlation of complete data is lower than both training and test data. We expect it to be lower than training correlation as the training data is seen by the model and subsequently should result in the highest correlation; however, that usually does not occur for the test data. The reason behind it is that the model trained on either SNRs is able to successfully predict the quality for the other SNR, but when applied on the complete data set, it is not as successful in following the trend between two different SNRs. It is worth mentioning that in all cases in this test, MARS has either a better or similar performance as that of SVR.

Table 5.8 reports the performance of three quality measures for partitioning based on the device under test (*Test 4*). For each case, the recordings made by four HAs are used for training and the recordings made by the other three HAs are used for testing. This means that 57% of data was used for training and 43% was used for testing. LCQA-HA performance is poor compared to the other two objective measures in *Test 4*, indicating that this measure does not generalize to this type of partitioning. Z-scores suggest that the difference between the performance of FBE-HL and PLP-HL is statistically significant. There is no statistically significant difference between MARS and SVR applied on FBE-HL in this test. In Figure 5.3, the relationship between the three nonintrusive measures and the cepstrum correlation feature extracted in HASQI for *Test 4* is plotted. FBE-HL fitted by SVR has the highest correlation and lowest standard deviation of error.

Table 5.7: Performance of SVR and MARS models on three objective quality measures: *Test 3*

Mapping Approach	**Objective Measure**	**Training Data**	ρ		
			Train	**Test**	**Complete**
SVR	LCQA-HA	*SNR0*	0.78	0.55	0.63
		SNR5	0.79	0.51	0.32
	FBE-HL	*SNR0*	0.96	0.96	0.86
		SNR5	0.98	0.93	0.85
	PLP-HL	*SNR0*	0.84	0.87	0.74
		SNR5	0.89	0.83	0.74
MARS	LCQA-HA	*SNR0*	0.81	0.80	0.68
		SNR5	0.84	0.76	0.64
	FBE-HL	*SNR0*	0.95	0.94	0.84
		SNR5	0.98	0.86	0.83
	PLP-HL	*SNR0*	0.89	0.90	0.79
		SNR5	0.95	0.87	0.75

Table 5.8: Performance of SVR and MARS models on three objective quality measures: *Test 4*

Mapping Approach	**Objective Measure**	**Training Data**	ρ		
			Train	**Test**	**Complete**
SVR	LCQA-HA	*HAOPSS*	0.83	0.27	0.56
		HASSRW	0.83	0.27	0.47
	FBE-HL	*HAOPSS*	0.96	0.75	0.83
		HASSRW	0.96	0.86	0.92
	PLP-HL	*HAOPSS*	0.75	0.79	0.77
		HASSRW	0.83	0.69	0.76
MARS	LCQA-HA	*HAOPSS*	0.67	0.40	0.52
		HASSRW	0.69	0.36	0.39
	FBE-HL	*HAOPSS*	0.94	0.84	0.84
		HASSRW	0.96	0.84	0.89
	PLP-HL	*HAOPSS*	0.80	0.75	0.77
		HASSRW	0.88	0.71	0.79

The performance of three measures for *Test 5* is reported in Table 5.9. In this test, the stimuli for two HL profiles are used for training and the other HL profile is used for testing. This implies that training on 67% of data and testing on 33% of data. LCQA-HA performs poorly in *Test 5* too. PLP-HL, while performing good in two cases, fails in predicting the quality at the other two cases. This means that the features extracted in PLP-HL are not robust (do not contain relevant information) regarding the type of HL. In fact, training the PLP-HL feature set with HA recordings for the S3 and N2 audiograms, and testing the trained function on the N4 recordings resulted in a negative correlation with the corresponding HASQI-CC scores. Thus, in this case, the PLP-HL coefficients were inadequate in capturing the necessary details of the auditory spectrum after accounting for the HL specified by the audiogram. FBE-HL, however, is robust in that regard and performs as good when trained on different partitions.

In Figure 5.4 (c., d., and e.), correlation coefficients are used to compare the performance of the three measures fitted by MARS and SVR and performed on *Test 3*, *Test 4*, and *Test 5*. There is no statistically significant difference between SVR and MARS applied on PLP-HL and FBE-HL for *Test 3* and *Test 4*. In *Test 5*, the differences between MARS and SVR on FBE-HL and PLP-HL are significant except for S3N2 test.

In HIRC database, the highest correlations are obtained with CC raw feature and as a result those correlations are reported in this work. Averaged spectral entropy and standard deviation of cepstral coefficients with correlation of 0.73 and 0.65, respectively; are the two most significant features in PLP-HL objective measure. After being fitted to HASQI-CC, the correlations vary from 0 in *Test 6* to 0.79 in *Test 5*.

Averaged iSNR with correlation coefficient of 0.25 and kurtosisof (d/dt (spectral dynamics)) with correlation coefficient of -0.25 score higher among all features extracted in LCQA-HA measure. After being fitted to HASQI-CC, the correlations vary from 0.10 in *Test 6* to 0.68 in *Test 3*.

Standard deviation of energies in 9th and 10th frequency bands in FBE-HL metric have correlations of 0.77 and 0.76. After being fitted to MOS, correlations vary from 0.63 in *Test 6* to 0.92 in *Test 4*.

Table 5.9: Performance of SVR and MARS models on three objective quality measures: *Test 5*

Mapping Approach	**Objective Measure**	**Training Data**	ρ		
			Train	**Test**	**Complete**
SVR	LCQA-HA	*N4S3*	0.87	0.41	0.30
		S3N2	0.83	0.64	0.58
	FBE-HL	*N4S3*	0.94	0.55	0.77
		S3N2	0.96	0.82	0.76
	PLP-HL	*N4S3*	0.72	0.24	0.77
		S3N2	0.75	0.13	-0.17
MARS	LCQA-HA	*N4S3*	0.82	0.43	0.43
		S3N2	0.77	0.31	0.58
	FBE-HL	*N4S3*	0.93	0.60	0.86
		S3N2	0.92	0.57	0.78
	PLP-HL	*N4S3*	0.81	-0.45	0.56
		S3N2	0.83	0.61	0.79

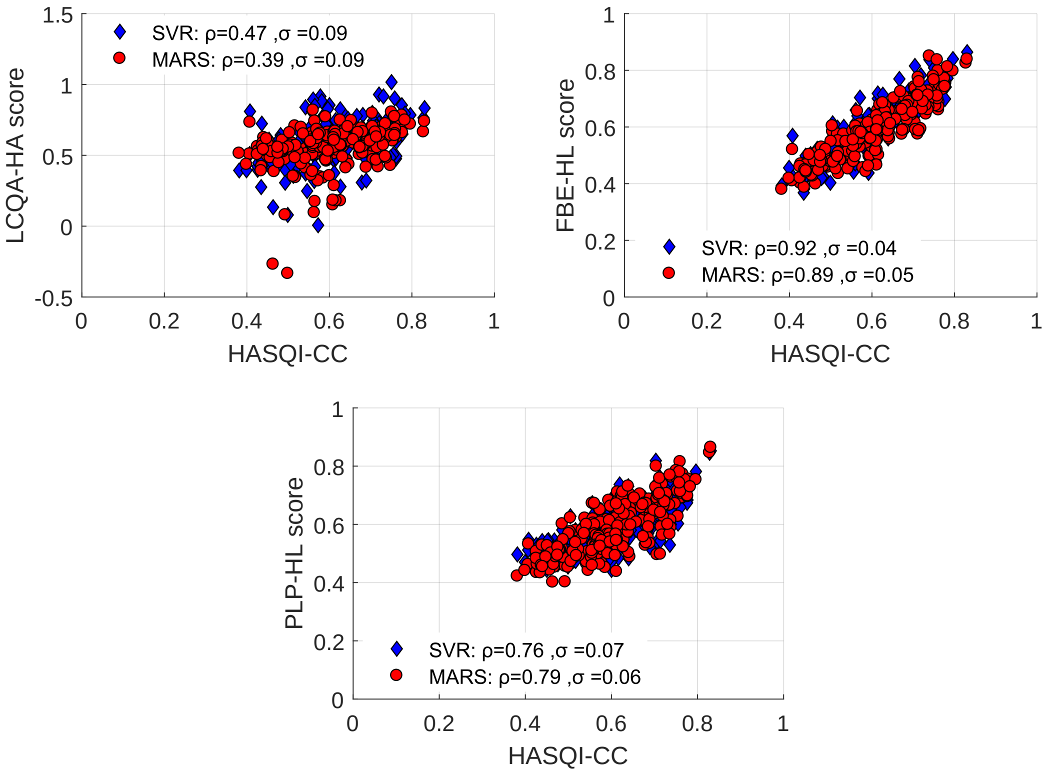

Figure 5.3: Relationship between HASQI-CC and predicted quality scores for LCQA-HA, FBE-HL, and PLP-HL measures fitted by SVR and MARS in *Test 4*: HIRC database

5.4.3 Cross-Database Validation

Finally, an experiment was conducted to evaluate the generalizability of the feature mapping procedure, wherein the prediction accuracy of LCQA-HA, FBE-HL, and PLP-HL features was evaluated when they were trained on binaural database and tested on HIRC database, and vice versa. The results are shown in Table 5.10.

LCQA-HA performs better on the binaural database compared to the HIRC database. The reason could be that in HIRC database there is more variability that is not picked up by LCQA-HA which utilizes a more compact modeling of spectrum. The variability in the HIRC database comes from using more HAs from different manufactures. Hence, the HAs have different bandwidths and different amplitude compression rationales which could potentially change the frame-to-frame spectral dynamics. Plot f. in Figure 5.4 shows the performance in terms of correlation coefficients for the cross-database test (*Test 6*). The

Table 5.10: Performance of SVR and MARS models on three objective quality measures: *Test 6*

Mapping Approach	Objective Measure	Training Data	ρ	
			Train	**Test**
SVR	LCQA-HA	*Binaural*	0.95	0.16
		HIRC	0.73	0.10
	FBE-HL	*Binaural*	0.95	0.76
		HIRC	0.95	0.85
	PLP-HL	*Binaural*	0.91	0.19
		HIRC	0.78	0.66
MARS	LCQA-HA	*Binaural*	0.96	0.14
		HIRC	0.74	0.10
	FBE-HL	*Binaural*	0.96	0.80
		HIRC	0.93	0.63
	PLP-HL	*Binaural*	0.95	0.00
		HIRC	0.83	-0.72

lowest correlations after optimization occur in *6* for all three measures which is expected as one database is used for training and a different database is used for testing the model.

LCQA-HA fails in predicting the quality of speech when it is trained on either one of the databases and tested on the other database. PLP-HL, while resulting in an average performance when trained on HIRC and tested on binaural database, fails in predicting the quality of HIRC database. In one case, when PLP-HL is fitted to HIRC database, it results in a negative correlation after being tested on binaural data.

The differences between the two mapping techniques in the cross-database validation for FBE-HL and PLP-HL are significant. It can be seen from Table 5.10 that a model trained on HIRC database results in a negative correlation with MOS when tested on binaural database for PLP-HL methodology. After looking back at the data, the reason appears to be in the fact that the features occupy different range. In fact when re-normalization is performed before applying MARS the correlation of -0.72 changed to 0.42. This suggests that MARS is sensitive to the data scaling and a proper scaling has to be performed. This becomes more crucial when MARS is trained on one database and tested on another database.

As it can be seen from the scatter plots in Figure 5.5, the predicted scores in plots b., c., and d. are outside the range for MOS in binaural database and HASQI-CC in HIRC database. In order to bring the predicted data into range two sets of scaling is done on input data. In method *A*, features are normalized by the maximum of absolute value of the data following the standardization in SVR. In method *B*, the standardization in SVR is disabled but the data is still normalized in the same way as the first method. The results are shown in Figure

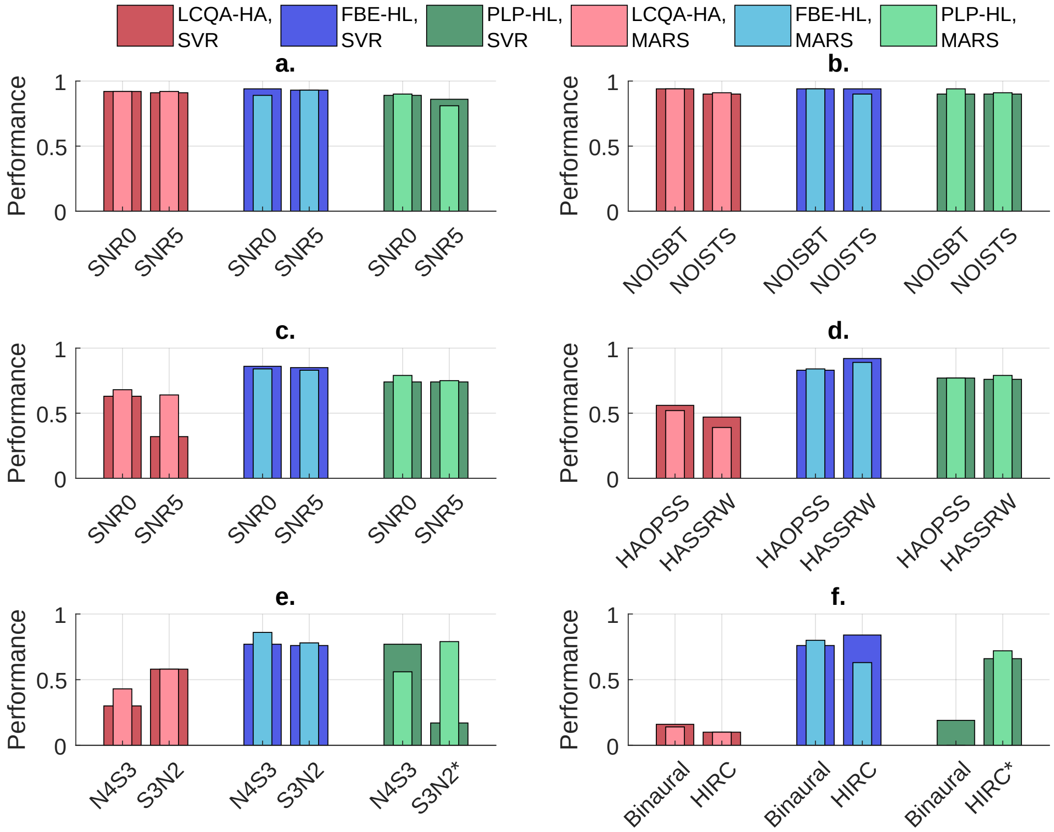

Figure 5.4: Absolute value of correlation coefficients of LCQA-HA, FBE-HL, and PLP-HL with different partitioning of data into training and test sets: a. *Test 1*, b. *Test 2*, c. *Test 3*, d. *Test 4*, e. *Test 5*, and f. *Test 6* (** Absolute value of the correlation coefficients reported in Tables 5.10 and 5.9 are shown.*)

5.6.

Table 5.11 lists the performance of ModA, SRMR-HA and HASQI on binaural database. Since the subjective data for HIRC database is unavailable, performance of HASQI can not be evaluated. ModA shows a relatively high performance for binaural database, however, it has a poor correlation with HASQI-CC for HIRC database. Correlation of SRMR-HA measure with MOS is 0.67, and 0.41 with HASQI-CC for HIRC database. This indicates that SRMR-HA does not generalize to HIRC database.

Table 5.11: Correlation coefficient and standard deviation of error for different objective quality metrics for binaural and HIRC databases

Database	Objective Measure	ρ	σ
Binaural	ModA	0.81	6.43
	SRMR-HA	0.67	8.13
	HASQI-CC	0.86	5.59
	HASQI-nonlinear	0.83	6.11
	HASQI	0.83	6.11
HIRC	ModA	0.51	0.09
	SRMR-HA	0.41	0.09

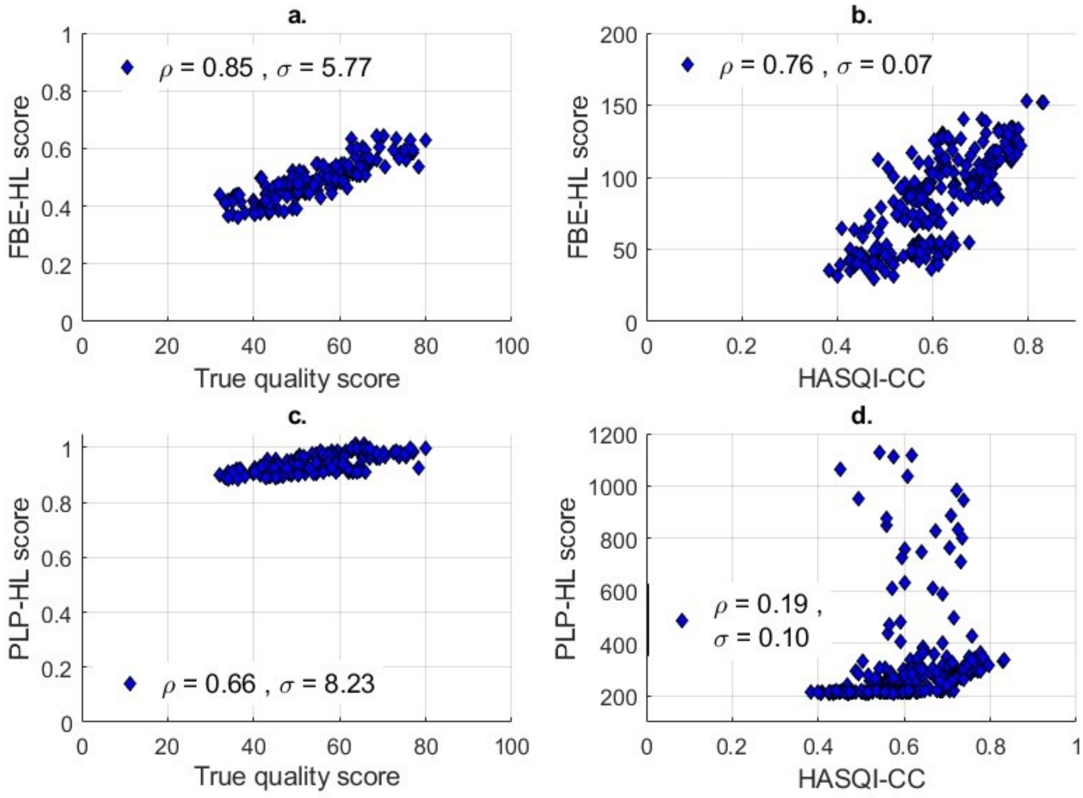

Figure 5.5: Relationship between MOS and HASQI-CC for FBE-HL, and PLP-HL scores trained by SVR on a. HIRC, b. binaural, c. HIRC, and d. binaural: *Test 6*

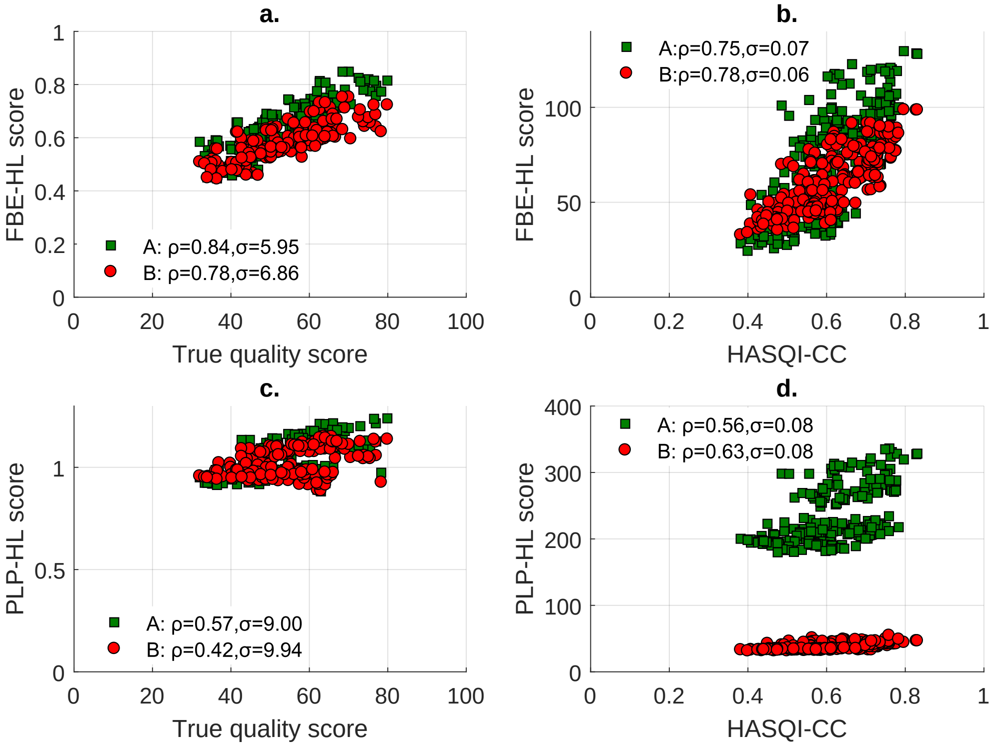

Figure 5.6: Relationship between MOS and HASQI-CC for re-normalized FBE-HL, and PLP-HL scores trained by SVR on a. HIRC, b. binaural, c. HIRC, and d. binaural: *Test 6*

5.5 Summary

This chapter was on validating the proposed objective speech quality metrics on two HI databases. Three existing nonintrusive metrics were also investigated on both databases. HASQI was applied as an intrusive metric. LCQA-HA along with the proposed metrics were mapped to MOS and HASQI-CC and their performance was analyzed. Different tests are performed to test generalizability of the proposed metrics. The results suggest that FBE-HL is robust for the different tests performed in this chapter and is a good candidate for speech quality assessment in HI applications.

Chapter 6

Wireless Remote Microphone Database

6.1 Introduction

Communication in demanding environments (e.g., noisy and/or reverberant environments) is a significant challenge faced by HI individuals [2]. HAs equipped with directional microphone processing are the most common treatment solutions to address this problem [36]. In general, directional microphones result in 2-5 dB improvement in speech reception thresholds (SRTs) in favourable environmental settings, but this benefit decreases when reverberation and distance from the source are factored in [36]. In such challenging acoustic environments, a wireless RM system can significantly enhance the speech perception abilities of HI listeners. RMs offer a substantial benefit to HI listeners in challenging environments with greater levels of background noise and reverberation. Contemporary RMs differ in terms of their microphone configuration (omnidirectional vs. directional), wireless communication protocols (FM vs. adaptive FM vs, digital RF), and additional signal processing (adaptive gain, noise reduction etc.). Furthermore, the coupling of RMs to personal HAs may result in unwanted changes to the gain/ output in HAs and may lead to unwanted distortions.

In this chapter, a new database is created from four RMs from different manufacturers interfacing a behind the ear (BTE) HA. This database is subsequently used to test generalizability of the proposed metrics. The results are compared with two existing nonintrusive metrics and HASQI.

6.2 Wireless Remote Microphone Database

6.2.1 Remote Microphones and Hearing Aid

Four RMs from three different manufacturers were assessed in this paper and their brief technical characteristics are given below:

1. Comfort Audio [118] digital microphone DM10 and the micro receiver DT10. This RM system utilizes a digital wireless communication protocol (Secure Stream Technology), supports a dynamic range of 60 dB, and an audio bandwidth of 100 Hz-7000 Hz. The microphone is omnidirectional.

2. Oticon Amigo-T31 and Amigo R2 receiver [119]. This RM system uses an adaptive FM protocol with automatic adjustment of FM emphasis based on background noise level (i.e., VoicePriority i [VPi]). It supports an audio bandwidth from 100 Hz to 8500 Hz and has a configurable omnidirectional or directional microphone.

3. Phonak EasyLink and MicroLink (MLxi) receiver [120]. This RM system uses adaptive (dynamic) FM technology and supports audio bandwidth from 100 Hz up to 7000 Hz.

4. Phonak Roger-Inspiro and MicroLink (MLxi) receiver [121]. This RM systems uses an adaptive digital wireless communication protocol in the 2.4 GHz band and a directional microphone. It supports an audio bandwidth of 100 Hz to 7300 Hz. This system also features adaptive gain control at the receiver depending on the background noise level, with the range of gain adaptation larger than the dynamic FM.

A commercially available behind the ear (BTE) HA (Unitron Quantum Pro S) was used for interfacing to all four RMs under test. This HA was programmed to match the DSL 5.0 adult prescriptive targets for the "N4" standard audiogram. All advanced signal processing features in the HA such as noise reduction, speech enhancement, and feedback cancellation were turned off. Fit to targets at soft (55 dB SPL), medium (65 dB SPL), and loud (75 dB SPL) input levels, and maximum power output (MPO) were verified in the Audioscan Verifit hearing aid test system. Once the HA fitting was verified, RM transparency was assessed using the Verifit system. The HA was connected to each RM receiver individually through the Direct Audio Input (DAI) and the HA was set to RM only (i.e., HA microphone off). Transparency criterion stipulates that equal inputs to the RMs under test must generate equal outputs from the HA, and that transparency is met if the average difference between HA only and RM only curves at 750, 1000 and 2000 Hz is $< \pm 2\ dB$. However a difference of $\pm 5\ dB$ was considered acceptable in a recently published work comparing different RMs. Figure 6.1 displays the frequency response curves obtained in Verifit, where the panel shows the long-term averaged spectra for 65 dB SPL speech input. The four RM responses are indicated by different colors in this figure, with the crosses denoting the DSL 5.0 targets.

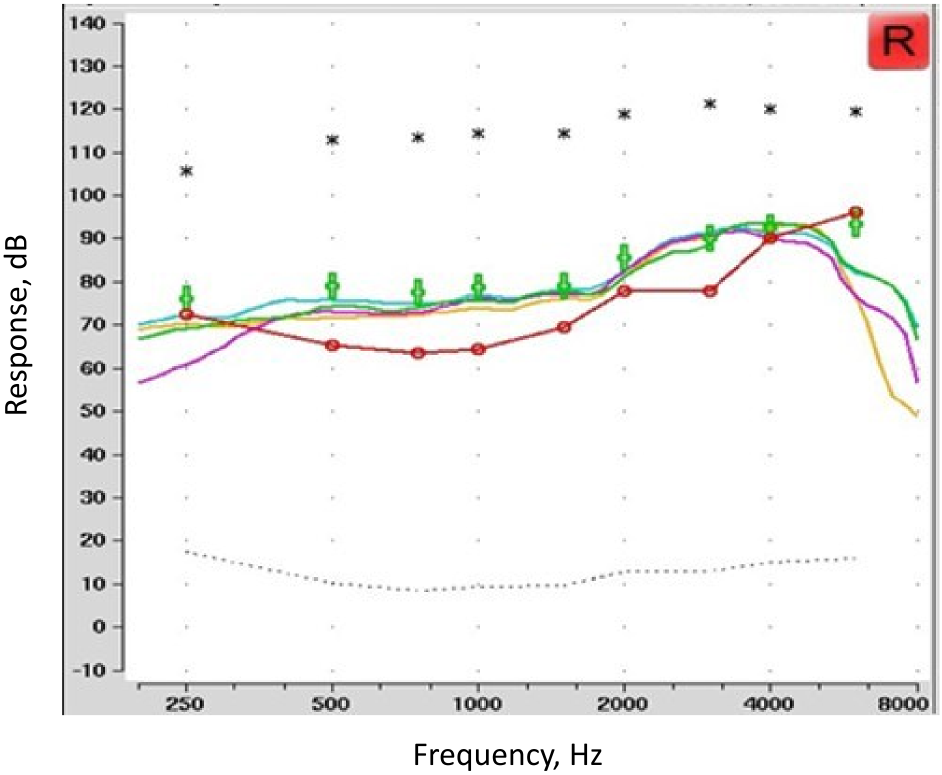

Figure 6.1: Frequency response of four RMs under test at 65 dB SPL speech input; Phonak EasyLink (orange), Phonak Roger-Inspiro (magneta), Oticon Amigo-T31 (cyan), and Comfort Audio DM10 (green).

It is evident from this Figure that RM frequency responses were similar to each other between 500 Hz - 4000 Hz at 65 dB SPL input. The average difference among the four RMs at 750, 1000, and 2000 Hz was $< \pm 2\ dB$, indicating that transparency was achieved. Even at an input level of 75 dB SPL, the average difference across the same frequencies was 5 dB, with the highest difference observed at 750 Hz.

6.2.2 Experimental setup and data collection

For electroacoustic measurements, the BTE HA + RM receiver assembly was connected to an ear mold simulator and placed on a HATS. This HATS served as the "listener". The RM transmitter was placed on a different HATS with a built-in mouth simulator, which served as the "talker" (see Figure 6.2). Speech stimuli (IEEE Harvard sentences) were presented through the mouth simulator and separate speaker(s) were used to present broadband background noise. The RM transmitters' microphone was placed 20 cm away from the centre of mouth simulator where the measured speech presentation level was 80 dBA. Recordings at the listener HATS in response to speech playback at talker HATS were collected for the following conditions: in quiet, and with uncorrelated noise at 0, and 10 dB SNR. The listener HATS was calibrated using the Bruel & Kjaer acoustic calibrator.

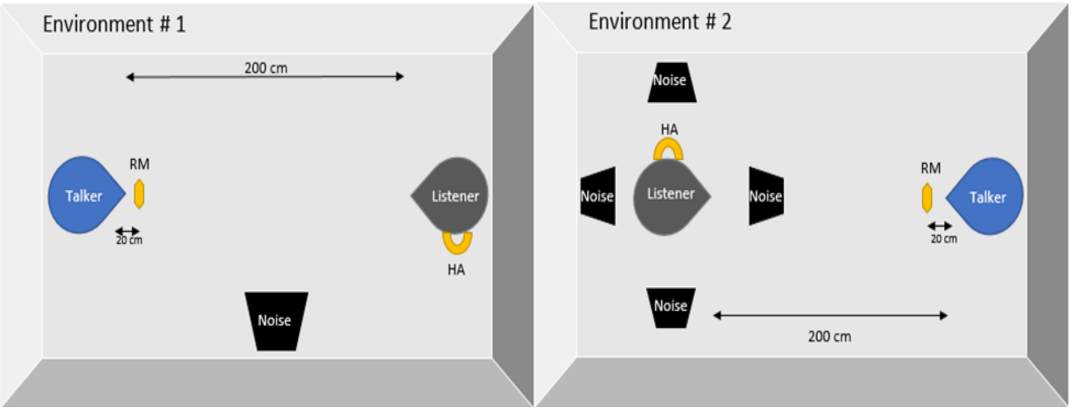

Figure 6.2: Experimental setup for RM evaluation in an acoustically benign room (left), and a harsh room (right).

The tests were performed in two different environments: an acoustically benign sound booth ("environment #1") with low reverberation ($RT_{60} = 0.1s$) and with only one single noise source; and an acoustically harsher reverberation chamber ("environment #2") with a higher degree of reverberation ($RT_{60} = 0.76s$) and surround noise. The speech was played back through the built-in mouth simulator and noise was played back through one speaker placed half the way and perpendicular to the connecting line in environment #1, and through four speakers at 0°, 90°, 180°, and 270° azimuth in environment #2, as shown in Figure 6.2. In both environments, the noise level (and hence the SNR) was measured at the centre of listener's head. All RM recordings were digitized at 16000 Hz sample rate and 16 bits/sample, and stored on a computer for offline analyses. The antialiasing filter in the A/D device ensures that no aliasing occurs when the recordings are digitized. As described in the previous section, some RMs have the option of selecting between omnidirectional and directional configuration for the transmitting microphone. In such cases, data was collected separately for the two different microphone configurations.

6.3 Methodology

ModA, SRMR-HA, LCQA-HA and HASQI are applied on RM database along with the two proposed metrics. In order to validate the proposed metrics and examine their robustness on RM database, the two models trained on binaural and HIRC databases are applied on RM database and the results are reported.

6.3.1 Feature Extraction

In order to apply PLP-HL feature set on RM database, the recordings are first high pass filtered and are re-scaled afterwards to reflect their absolute SPL. The HL profile is set to "N4" standard audiogram. The order of recursion in Levinson-Durbin method is set to 6 and the number of filters in gammatone filterbank is set to 32. Following this procedure, 12-dimensional feature vector is calculated through PLP-HL metric.

In order to extract LCQA-HA features, the same procedure as outlined in chapter 4 is followed. In order to compute FBE-HL, the recordings are re-scaled afterwards to reflect their absolute SPL. In order to investigate the effects of HL model on FBE performance two sets of measurements are performed. In order to apply the HL model, the weigheted spectrum is compressed and then the compressed spectrum is converted to dB SL. Finally, mean and variance of the energy in each of 32 frequency bands is calculated to give rise to 64 features.

SRMR-HA is the last nonintrusive metric applied on the recordings and the outputted index is captured for the entire database. HASQI requires a reference signal to predict the quality of signal under test. Before any comparison takes place, the reference signal should be normalized to have RMS value of 1 and then frequency shaped using REAGs. Then, the signal under test is high pass filtered and normalized and subsequently time-aligned with reference signal. These two signals along with the corresponding HL profile and the level of the input signal are inputted to HASQI version 2 algorithm.

6.3.2 Feature Mapping

The models trained on binaural and HIRC databases by SVR and MARS are tested on RM database. The results are reported and discussed in the following sections.

6.4 Results

6.4.1 Remote Microphone Performance Evaluation

The four RMs investigated in this study were randomly labeled RM_A, RM_B, RM_C, and RM_D. Figure 6.3 displays the spectrograms computed from a sample set of RM recordings in environment #1, which allow for gauging the frequency range and noise level. Figures 6.3(a) and 6.3(b) depict the spectrograms of RM_A recordings in speech in quiet and speech in noise (SNR = 0 dB) conditions respectively, while Figures 6.3(c) and 6.3(d) show the corresponding spectrograms for RM_B. It is evident from Figure 6.3(c) that the bandwidth of RM_B is limited to less than 6 kHz, contrary to its specifications. Moreover, a higher internal noise level in higher frequencies can be observed with RM_B recording in Figure 6.3(c). Figure 6.3(b) shows that the RM_A is more robust to the background noise than RM_B, mainly due to the directional microphone configuration at its transmitter.

Figure 6.4 displays the HASQI, and SRMR-HA values obtained from the RM recordings in both environments, with no background noise and in the presence of background noise at 10 dB and 0 dB SNRs. It must be noted here that HASQI values are normalized to a range of 0-1, while there was no normalization of the SRMR-HA values. For both metrics higher values indicate better quality.

As expected and evident in Figure 6.4, the objective metrics are lower with an increase in the background noise level and \or reverberation. Taking a closer look at the HASQI data, it can be seen that all RMs exhibit similar performance in quiet in environment #1. This condition is similar to the transparency verification condition in the test box, and implies that all RMs perform similarly when there is no background noise and low reverberation.

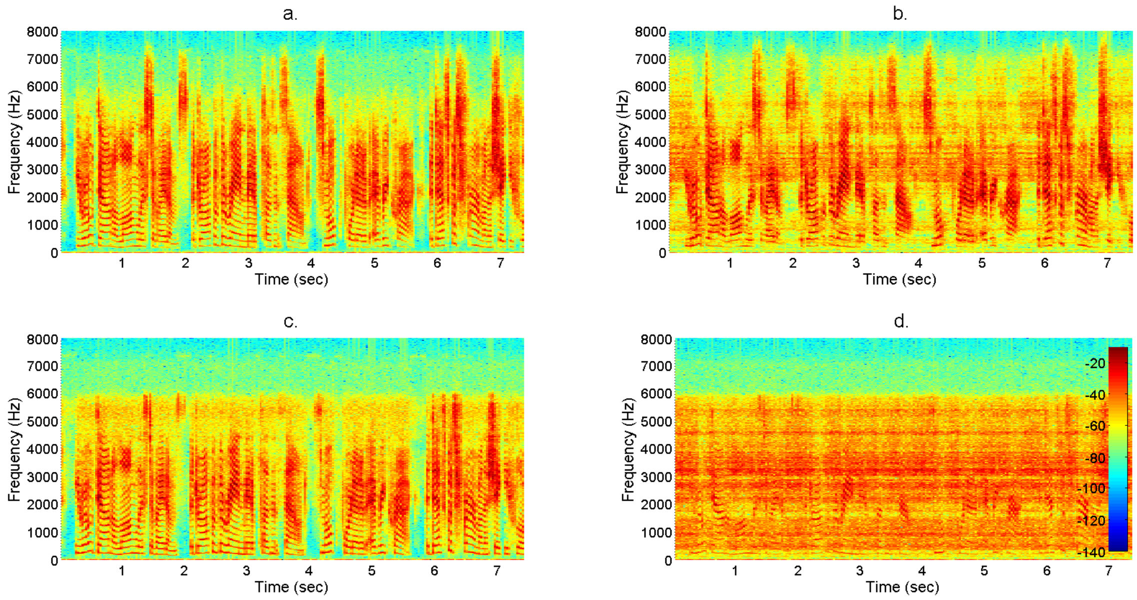

Figure 6.3: Spectrograms of the RM recordings in environment #1. (a) Speech in quiet for RM_A , (b) Speech in 0 dB SNR for RM_A, (c) Speech in quiet for RM_B , and (d) Speech in 0 dB SNR for RM_B.

However, differences do emerge in RM performance in quiet and in the presence of reverberation (Figure 6.4(b)) and in the presence of background noise in both environments. At SNR = 10 dB, RM_A has the best performance followed by RM_C utilizing a directional microphone. The performance gap between RM_A and the rest of the RMs widened at 0 dB SNR (Figures 6.4(a) and (b)). Several factors may have contributed to the comparatively better performance of RM_A, including the presence of directional microphone which not only reduces background noises but also partially attenuates reverberation components. In addition, RM_A incorporates an adaptive gain control strategy at the receiver. In general, HASQI and SRMR-HA values show a similar trend across noise and reverberation conditions. In fact, the correlation coefficients of SRMR-HA with HASQI was 0.71. The lower correlation coefficients exhibited by SRMR-HA are mainly due to the discrepancies in scores for RM_D, SRMR-HA ranks RM_D more favourably, especially in quiet conditions. In order to gain further insight into this, modulation spectrogram plots were obtained from RM_A and RM_D recordings and displayed in Figure 6.5. In these plots, the x-axis represents the center frequency of the modulation filterbank, the y-axis represents the center frequency of the gammatone filterbank, and the colors represent the relative modulation energy. Recall that the SRMR-HA computes the ratio of the averaged modulation energies between 4 to 18 Hz and 29 to 128 Hz modulation channels. It can be noticed from Figures 6.5(a) and

(c) that the RM_A recording in quiet has a pocket of energy distribution in the upper modulation frequency region, while the RM_D recording is devoid of it. As such, the SRMR-HA resulted in a higher score for RM_D for the speech in quiet condition. At an SNR of 10 dB (Figures 6.5(b) and (d)), it can be seen that the modulation spectrogram of RM_A recording was relatively unchanged (thus highlighting the robustness of RM_A), while that of RM_D recording was significantly affected.

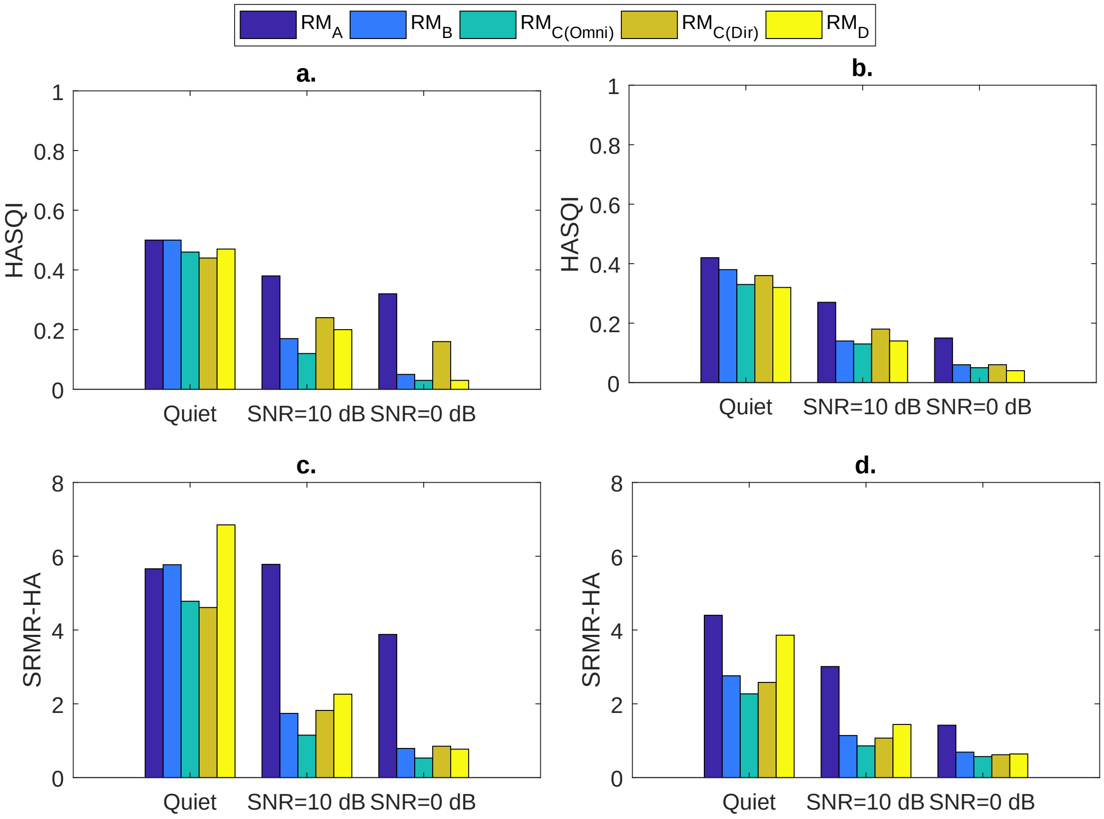

Figure 6.4: Objective speech quality (HASQI and SRMR-HA) values for different RMs across different noise conditions and environments; (a) HASQI in environment #1, (b) HASQI in environment #2, (c) SRMR-HA in environment #1, and (d) SRMR-HA in environment #2.

6.4.2 Feature Set Validation

In the same manner as chapter 5, the model trained on either binaural or HIRC database is tested on RM database and results are reported in Table 6.1. From the results it is concluded that FBE-HL is robust when optimized by both SVR and MARS methods. While PLP-HL shows good performance for SVR, it is not robust for MARS when the model

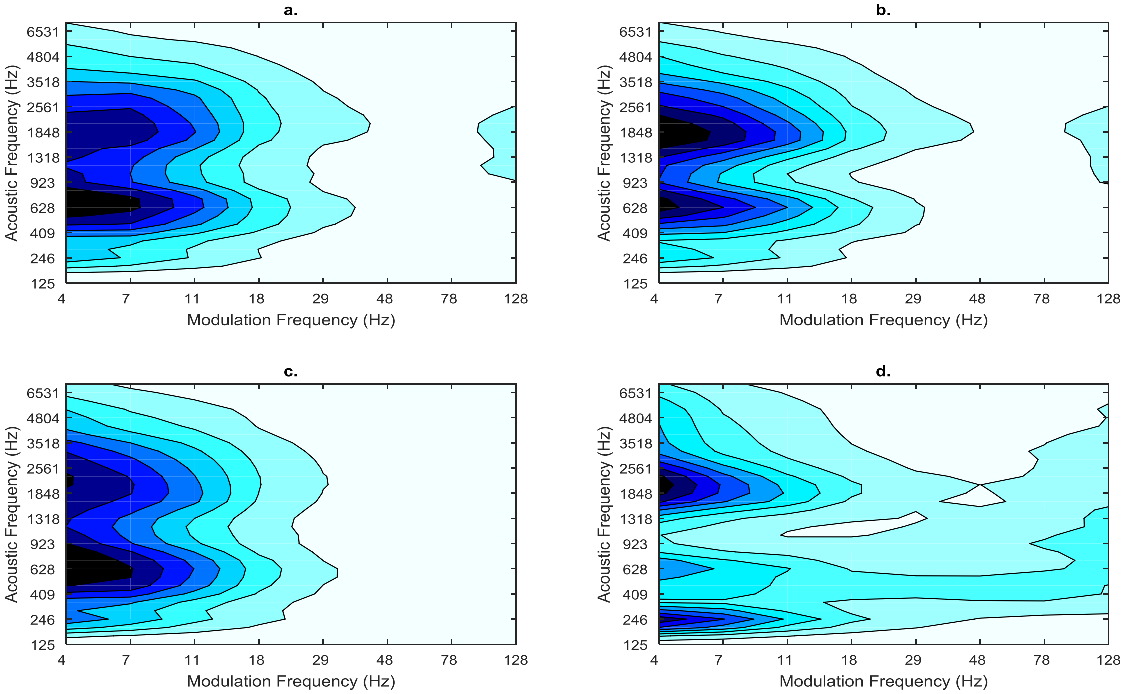

Figure 6.5: Modulation spectrograms computed from RM_A and RM_D recordings in environment #1. (a) RM_A in quiet, (b) RM_A in SNR = 10 dB , (c) RM_D in quiet , (d) RM_D = 10 dB SNR.

is trained on HIRC database. LCQA-HA, however, shows poor performance with both learning approaches.

As it can be seen from the scatter plots in Figure 6.6, the predicted scores in plots b., c., and d. are outside the range for MOS in binaural database and HASQI-CC in HIRC database. In order to bring the predicted data into range, two sets of scaling is done on the input data. In method *A*, features are normalized by the maximum of absolute value of the data following standardization in SVR. In method *B*, the standardization in SVR is disabled but the data is still normalized in the same way as the first method. The results are shown in the following scatter plots.

Finally, performance of ModA, SRMR-HA and HASQI on RM database is reported in Table 6.2. It should be highlighted that the correlation coefficients reported here are between the objective measures and HASQI-CC raw feature.

Table 6.1: Performance of SVR and MARS models trained on binaural and HIRC data sets for the objective quality measures

Mapping Approach	Objective Measure	Training Data	ρ	
			Train	Test
SVR	LCQA-HA	*Binaural*	0.95	0.10
		HIRC	0.73	0.38
	FBE-HL	*Binaural*	0.95	0.94
		HIRC	0.95	0.90
	PLP-HL	*Binaural*	0.91	0.94
		HIRC	0.78	0.91
MARS	LCQA-HA	*Binaural*	0.96	-0.10
		HIRC	0.74	0.51
	FBE-HL	*Binaural*	0.96	0.90
		HIRC	0.93	0.66
	PLP-HL	*Binaural*	0.95	0.89
		HIRC	0.83	-0.64

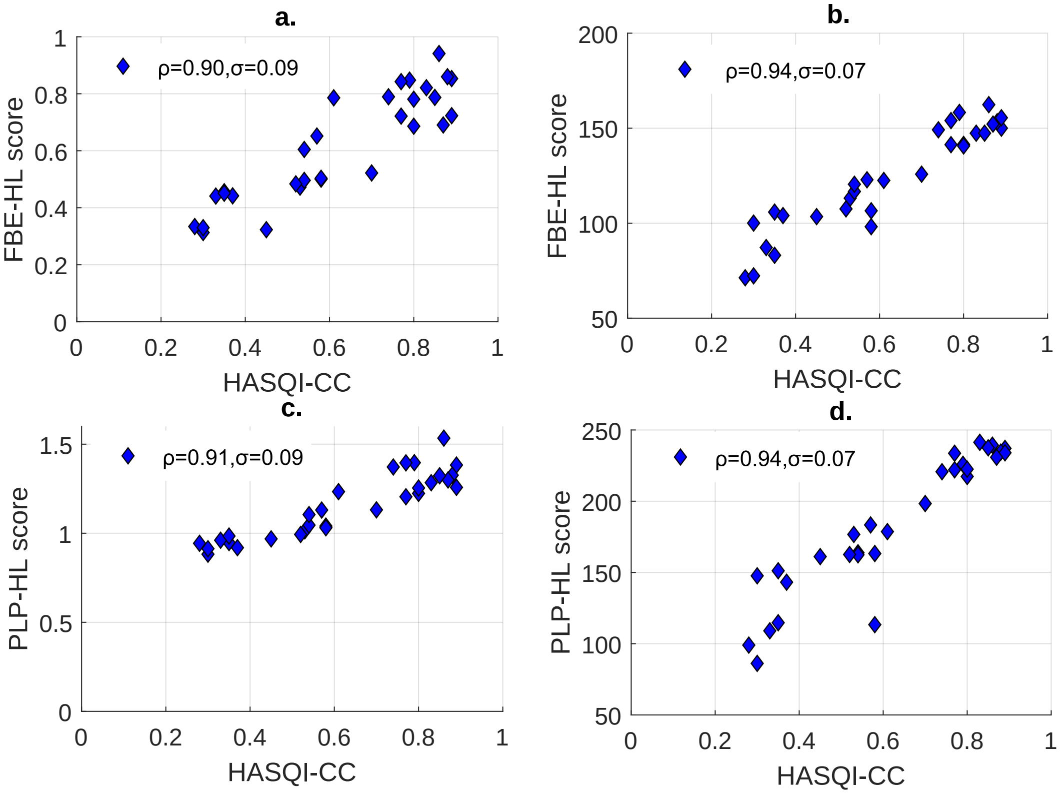

Figure 6.6: Relationship between HASQI-CC and predicted quality scores for FBE-HL and PLP-HL measures trained by SVR trained on a. HIRC, b. binaural, c. HIRC, and d. binaural

Table 6.2: Correlation coefficient and standard deviation of error for different objective quality metrics RM database

Objective Measure	ρ	σ
ModA	0.96	0.06
SRMR-HA	0.71	0.15

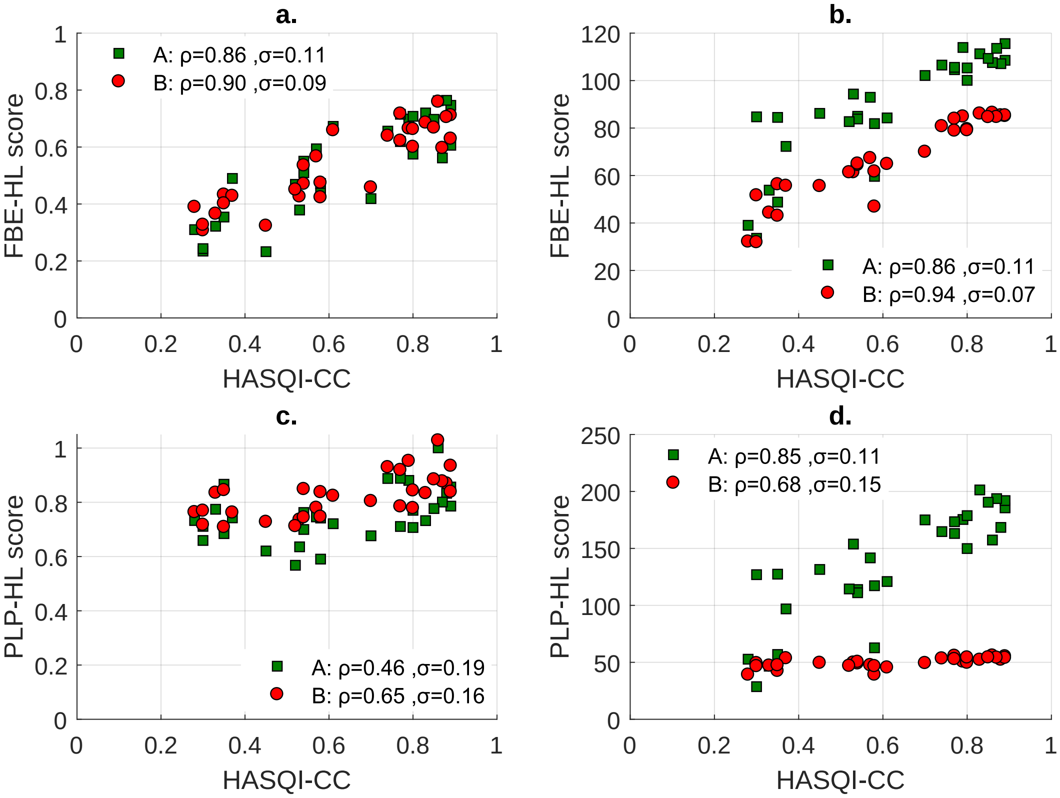

Figure 6.7: Relationship between HASQI-CC and predicted quality scores for re-normalized FBE-HL and PLP-HL measures trained by SVR on a. HIRC, b. binaural, c. HIRC, and d. binaural

6.5 Discussion

6.5.1 Remote Microphone Systems

Wireless RMs are an attractive assistive listening device option for HI listeners in challenging acoustical environments. Behavioural studies have shown that these inter-device differences do lead to performance differences, with some RMs performing better than others. In this chapter the relative performance of four different RMs was benchmarked by HASQI and SRMR-HA metrics. While these metrics have not been directly validated with behavioural data collected with RMs, it must be noted that all three metrics base their computation on the signal captured at the tympanic membrane (i.e., HA output). Thus any distortion or enhancement introduced by the RM will reflect in the HA output and therefore will have a corresponding influence on the calculated score. The metrics showed the expected trend that the RM performance degrades with an increase in background noise level and reverberation. What is noteworthy is the differential degradation in RM performance across SNRs and reverberation conditions. This is despite the fact that all RMs were verified to be transparent, suggesting that transparency verification alone is inadequate for characterizing RM performance.

It is clear from Figure 6.4 that RM_A has a significantly higher HASQI and SRMR-HA scores in both environments at 0 dB and 10 dB SNRs. In addition, internal noise measurements have shown that RM_A had the lowest noise floor among the four RM systems. RM_A has the following salient features: 1) proprietary digital wireless communication protocol in the 2.4 GHz band employing time and frequency diversity, 2) array microphone at the transmitter, and 3) additional signal processing at the receiver for automatic gain control in response to background noise level. Previous studies by Thibodeau [50] and Wolfe [122] have shown this RM to provide significantly better speech recognition in noise by HI listeners, in comparison to other RM technologies. Thus, the electroacoustic data presented in this chapter is aligned with the published behavioural data for this RM.

6.5.2 Feature Set Validation

In the second study, the proposed nonintrusive metrics are validated on the collected data through RMs. The highest correlations for the extracted features are obtained with HASQI-CC raw feature and as a result those correlations are reported in this work.

Standard deviation of PLP-HL coefficients and cepstral coefficients have the highest correlations in PLP-HL measure with 0.92 and 0.95 values; respectively. After optimization, the highest correlation is 0.94 for the model trained by SVR on binaural data and the lowest is -0.64 with the model trained by MARS on HIRC database.

The highest correlations in LCQA-HA measure occur for iSNR (r = 0.87) and averaged spectral centroid (r = -0.75). After the trained models on binaural and HIRC databases were tested on this database, the correlations vary from -0.10 to 0.51 indicating that models trained on LCQA-HA are not generalizable to RM data.

Standard deviation of energies in 21$^{\text{st}}$, 22$^{\text{nd}}$, and 23$^{\text{rd}}$ frequency bands in FBE-HL metric have correlations of 0.93. After being fitted to MOS, correlations vary from 0.66 to 0.94.

Z-scores suggest that the difference between the performance of FBE-HL and PLP-HL trained by either SVR or MARS on binaural database is statistically significant to models trained on HIRC database. The is no statistically significant difference between MARS and SVR applied on FBE-HL in this test.

From Figure 6.6, it is evident that although both metrics perform well on RM data, the predicted quality is outside range (except for FBE-HL trained on HIRC database). The reason is that the features are not occupying the same range. Plot d. in Figure 6.7 indicates that re-normalization brings PLP-HL scores into range but its performance decreases; however, in d. where method *A* is used, the predicted data is still outside range and while method *B* brings the data into range its performance is poor. In FBE-HL case, however, when plot a. of Figure 6.3 and 6.4 are compared, the predicted data is always inside range and FBE-HL does not fail when the choice of re-normalization is changed. Comparing plots b. in both figures shows that method *B* is successful in predicting quality data that are within the acceptable range with good performance.

Table 6.2 lists the performance of ModA, SRMR-HA on RM database. Since the subjective data is unavailable, performance of HASQI can not be evaluated.

6.6 Summary

In this chapter the relative performance of four different RMs was benchmarked by HASQI and SRMR-HA metrics. The results suggest that the devices performance degrades with increasing background noise. Also transparency alone is not adequate for characterizing RM performance.

As the second study, the proposed objective speech quality metrics were validated on the RM database. Three existing nonintrusive metrics were also investigated on the same database. HASQI was applied as an intrusive metric. LCQA-HA along with the proposed metrics were mapped to HASQI-CC and their performance was analyzed. Re-normalization was done to bring the predicted data into the range and the outcome was discussed. After re-normalization, FBE-HL has a correlation of 0.90 and 0.94 when tested on binaural and HIRC databases, respectively. PLP-HL scores 0.65 or 0.68 for the same tests. The results suggest that features extracted through FBE-HL methodology are generalizable to the RM database.

Chapter 7

Contributions and Future Works

In this chapter the contributions of this work are given. Recommendations for the future work are also included.

7.1 Contributions

- Two methods for feature extraction from distorted signals were introduced. PLP technique was selected as the base model for feature extraction; however, PLP does not take into account the effects of HL on auditory system. So, it was modified to incorporate the effects of sensorineural HL in a manner similar to HASQI. The loss of spectral resolution, elevated hearing thresholds, and abnormal growth of loudness that are side-effects of hearing loss were incorporated in PLP. These modifications makes the proposed methodology for feature extraction suitable for both NH and HI applications. In order to extract features, PLP-HL model coefficients are computed by Levinson-Durbin recursion and their statistical properties are found. The statistical properties along with the absolute value of skewness of the two sets of coefficients resulted in 10 features. By adding the averaged estimated prediction residual error power and spectral entropy, the final number of features in the PLP-HL feature set was 12. FBE-HL feature set is derived by calculating the gammatone filterbank energies. The mean and variance of the frame energies in each of the 32 channels were calculated, which gave rise to a 64 FBE-HL feature set.

- Three approaches were investigated for mapping the proposed feature sets and the feature set introduced in LCQA-HA to the subjective scores. Two-step dimensionality reduction followed by linear regression, SVR machine, and MARS are applied on randomly selected training sets and their performance is averaged on test sets. SVR and MARS are selected to train models on various portions of a NH database and test those models on the remaining data. The NH database is collected by enhancing noisy signals using seven NR algorithms. Four tests are performed. In *Test* 1 the data is partitioned based on the sentences spoken. In *Test* 2 partitioning is performed with respect to the applied NR algorithm while in *Test* 3 different SNRs are used for data partitioning. Finally in *Test* 4, the type of noise is used for selecting training and test sets. ModA, SRMR, and HASQI were also applied on the same database. The results show that FBE-HL feature set fitted by SVR has a better performance compared to PLP-HL and LCQA-HA and outperforms modA and SRMR as nonintrusive metrics. HASQI, however, resulted in better performance among all metrics examined in this study.

- The same procedure for mapping was performed on the proposed and LCQA-HA feature sets computed from two HI databases. The first database comprised of bilateral HA recordings in real environments while the second database consisted of unilateral HA recordings recorded in desktop HA test box. SNR levels and noise types were used to partition the binaural database. The second HI database was partitioned based on SNR levels, HA models, and the audiograms used for fitting the HAs. Performance of ModA, SRMR-HA, and HASQI was investigated on binaural database. However, since for the second database the CC feature from HASQI was used to serve as subjective scores, only the performance of modA and SRMR-HA was evaluated.

- An experiment was conducted to evaluate the generalizability of the feature mapping procedure, wherein the prediction accuracy of LCQA-HA, FBE-HL, and PLP-HL features was evaluated when they were trained on binaural database and tested on HIRC database, and vice versa. FBE-HL reported similar accuracy results for the cross-database validation, despite the differences in data collection methodology across the two databases. It was concluded that PLP-HL and LCQA-HA feature sets are incapable of generalizing across HA speech quality databases collected through different modalities.

- The need for proper feature normalization for cross-database validation when there are methodological differences in the collection of database content was discussed. Three methods for normalization were examined and the results were reported. Through this investigation, it is shown that FBE-HL is the more robust and generalizable feature set for HA speech quality applications.

- A new database consisted of recordings through four RMs paired with a HA was created. The recordings were made in real environments with the devices placed on HATS. HASQI and SRMR-HA were used to benchmark the devices. The results showed that the RM with directional microphones resulted in a better performance.

- The performance of the two proposed and LCQA-HA feature sets trained on the two HI databases was tested on the RM database. Again, FBE-HL feature set trained on either databases by SVR machine resulted in high correlation coefficients with the CC feature in HASQI which was used in place for subjective scores.

7.2 Study Limitations and Future Work

Based on the work presented in this thesis a number of recommendation exist for future work:

- This study focused on predicting quality of speech. It is encouraged to investigate the validity of the proposed metrics on music quality. The performance of HASQI in predicting music quality has been investigated in [123]. Hearing Aid Audio Quality Index (HAAQI) [124] has recently been introduced as an intrusive measure for music and audio quality assessment. The performance of HAAQI was examined on the same database as the one used in [123]. The data was captured using a simulated database. A study should be performed on real HAs to validate HASQI, HAAQI, and the proposed nonintrusive measure.

- The proposed metrics were validated on NH database with a focus on wideband noisy stimuli enhanced by different NR algorithms. This study could be extended to include other types of distortions and algorithms for NH applications like EC and speech codecs.

- The HI databases used in this study included three specific hearing aid features, namely NR, directionality, bilateral communication and RM systems. The techniques developed could be extended to the study of additional hearing aid features such as feedback cancellation, frequency compression etc.

- FBE-HL fitted by SVR machine resulted in the better performance for the majority of the test cases. Although SVR does not perform dimensionality reduction, different reduced feature sets (through PCA or MARS) were mapped by SVR but resulted in lower correlations with true quality scores. Looking at the individual correlations of FBEs for individual audiograms may result in more insights into a better approach for dimensionality reduction.

- Linear regression, SVR machine and MARS were investigated on the proposed methodology; however, other mapping techniques such as neural networks, CART, and GMM classifier could be investigated too.

- For future work, it is suggested to extend the study of the existing and proposed quality measures for performance evaluation of a wider range of RM devices and HAs.

- It is encouraged to study whether including the other effects of HL in the auditory system (e.g. decreased temporal resolution) into the proposed model would increase its performance in quality prediction.

Bibliography

[1] S. Launer, J. A. Zakis, and B. C. J. Moore, "Chapter 4: Hearing Aid Signal Processing," in *Hearing Aids*, 2016, vol. 56, no. December, pp. 93–130.

[2] S. Kochkin, "MarkeTrak VIII : Consumer Satisfaction with Hearing Aids Is Slowly Increasing," *Hearing*, vol. 63, no. 1, pp. 19–32, 2010.

[3] J. E. Preminger and D. J. Van Tasell, "Quantifying the Relation between Speech Quality and Speech Intelligibility," *Journal of Speech and Hearing Research*, vol. 38, pp. 714–725, 1995.

[4] ITU-T G. 722, "7 kHz Audio-Coding Within 64 kbit/s," 1988.

[5] "Wideband Speech Coding Standards and Applications," *VoiceAge White papers*, 2006.

[6] C. Beaugeant, M. Schönle, and I. Varga, "Challenges of 16 kHz in Acoustic Pre- and Post-Processing for Terminals," *IEEE Communications Magazine*, vol. 44, no. 5, pp. 98–104, 2006.

[7] P. Stelmachowicz, A. Pittman, B. Hoover, and D. Lewis, "Effect of Stimulus Bandwidth on the Perception of /S/ in Normal- and Hearing-Impaired Children and Adults," *Journal of the acoustical society of America*, vol. 110, no. 4, pp. 2183–90, 2001.

[8] S. Pennock and P. Hetherington, "Wideband Speech Communications: he Good, the Bad, and the Ugly," in *Audio Engineering Society*, Dearbon, Michigan, USA, 2009, pp. 109–117.

[9] T. Madhavi, B. H. Krishna, and L. P. Kanth, "A Novel Approach for Voice Quality Enhancement for VoIP Applications Using DSP Processor," *International Conference on Electrical, Electronics, and Optimization Techniques, ICEEOT 2016*, pp. 4946–4951, 2016.

[10] C. Faller and J. Chen, "Suppressing Acoustic Echo in a Spectral Envelope Space," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1048–1061, 2005.

[11] S. Zoican, "Noise Reduction and Echo Cancellation System," *6th International Conference on Signal Processing, 2002.*, vol. 2, no. I, pp. 1324–1327, 2002.

[12] R. Le Bouquin Jeannès, P. Scalart, G. Faucon, and C. Beaugeant, "Combined Noise and Echo Reduction in Hands-Free Systems: A Survey," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 808–820, 2001.

[13] E. Böhmler, J. Freudenberger, and S. Stenzel, "Combined Echo and Noise Reduction for Distributed Microphones," *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays, HSCMA'11*, no. 1, pp. 98–103, 2011.

[14] T. S. Wada and B.-h. Juang, "Acoustic Echo Cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 175–189, 2012.

[15] ITU-T, "Single-Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications," ITU-T P.563, Tech. Rep., 2004.

[16] ITU-T, "Mean Opinion Score (MOS) Terminology," Jun. 2006.

[17] ITU-T P. 835, "Subjective Test Methodology for Evaluating Speech Communication Systems That Include Noise Suppression Algorithm," 2003.

[18] "NIDCD Fact Sheet: Hearing Aids," National Institute on Deafness and Other Communication Disorders, Tech. Rep. Cic, 2008.

[19] H. B. Abrams and J. Kihm, "An Introduction to MarkeTrak IX: A New Baseline for the Hearing Aid Market," *Hearing Review*, vol. 22, no. 6, p. 16, 2015.

[20] L. Tran, H. Schepker, S. Doclo, H. Dam, and S. Nordholm, "Improved Practical Variable Step-Size Algorithm for Adaptive Feedback Control in Hearing Aids," *2016, 10th International Conference on Signal Processing and Communication Systems, ICSPCS 2016 - Proceedings*, no. 1, 2016.

[21] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Theoretical Analysis of Linearly Constrained Multi-Channel Wiener Filtering Algorithms for Combined Noise Reduction and Binaural Cue Preservation in Binaural Hearing Aids," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 12, pp. 2384–2397, 2015.

[22] P. Wang, B. Fan, Y. S. I, and G. Yang, "Optimized Realization of Wide Dynamic Range Compression based on DSP5535 Hearing Aid," *Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, pp. 1127–1131, 2016.

[23] H. Puder, E. Fischer, and J. Hain, "Optimized Directional Processing in Hearing Aids with Integrated Spatial Noise Reductian," *International Workshop for Acoustic Echo and Noise Control*, no. September, pp. 1–4, 2012.

[24] M.-y. Zhang, “Speech Recognition and Synthesis Algorithm for Digital Hearing Aids under Background Noise,” *International Conference on Information System and Artificial Intelligence Speech*, 2016.

[25] M. Jelinek and R. Salami, “Noise Reduction Method for Wideband Speech Coding,” *Proc. Eusipco*, pp. 1959–1962, 1959.

[26] J. D. Gordy and R. A. Goubran, “On the Perceptual Performance Limitations of Echo Cancellers in Wideband Telephony,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 33–42, 2006.

[27] B. N. M. Laska, R. A. Goubran, and M. Bolic, “Improved Proportionate Subband NLMS for Acoustic Echo Cancellation in Changing Environments,” *IEEE Signal Processing Letters*, vol. 15, pp. 337–340, 2008.

[28] H. Dillon, *Hearing Aids*. Boomerang Press, 2001.

[29] D. Hall, J. Llinas, G. M. Davis, and R. Crane, *Noise Reduction in Speech Applications*, G. M. Davis, Ed. CRC Press LLC, 2002.

[30] A. Schaub, Ed., *Digital Hearing Aids*. Thieme Medical Publishers, 2011.

[31] P. E. Souza, “Effects of Compression on Speech Acoustics, Intelligibility, and Sound Quality,” *Trends in Amplification*, vol. 6, no. 4, pp. 131–165, 2002.

[32] M. Valente, *Hearing Aids: Standards, Options, and Limitations*, 2nd ed. Thieme Medical Publishers, 2002.

[33] R. A. Bentler, “Effectiveness of Directional Microphones and Noise Reduction Schemes in Hearing Aids: a Systematic Review of the Evidence.” *Journal of the American Academy of Audiology*, vol. 16, pp. 473–484, 2005.

[34] E. Mackenzie and M. E. Lutman, “Speech Recognition and Comfort Using Hearing Instruments with Adaptive Directional Characteristics in Asymmetric Listening Conditions,” *Ear and hearing*, vol. 26, no. 6, pp. 669–679, Dec. 2005.

[35] A. M. Amlani and P. J. L. Rakerd, Brad, “Speech-Clarity Judgments of Hearing-Aid-Processed Speech in Noise: Differing Polar Patterns and Acoustic Environments,” *International Journal of Audiology*, vol. 45, no. 6, pp. 319–330, 2006.

[36] H. Dillon, *Hearing Aids*. Boomerang Press, 2012.

[37] “NIDCD Fact Sheet: Hearing Aids,” National Institute on Deafness and Other Communication Disorders, Tech. Rep. Cic, 2008.

[38] T. A. Ricketts and B. W. Y. Hornsby, “Sound Quality Measures for Speech in Noise through a Commercial Hearing Aid Implementing "Digital Noise Reduction",” *Journal of the American Academy of Audiology*, vol. 16, no. 5, pp. 270–277, May 2005.

[39] A. Pittman, “Age-Related Benefits of Digital Noise Reduction for Short-Term Word Learning in Children with Hearing Loss,” *Journal of Speech Language and Hearing Research*, vol. 54, no. 5, p. 1448, 2011.

[40] A. Sarampalis, S. Kalluri, B. Edwards, and E. Hafter, “Objective Measures of Listening Effort: Effects of Background Noise and Noise Reduction,” *Journal of Speech Language and Hearing Research*, vol. 52, no. 5, pp. 1230–1240, 2009.

[41] C. A. Quintino, M. F. C. G. Mondelli, and D. V. Ferrari, “Directivity and Noise Reduction in Hearing Aids: Speech Perception and Benefit,” *Brazilian Journal of Otorhinolaryngology*, vol. 76, no. 5, pp. 630–638, 2010.

[42] P. Stelmachowicz, D. Lewis, B. Hoover, K. Nishi, R. McCreery, and W. Woods, “Effects of Digital Noise Reduction on Speech Perception for Children with Hearing Loss,” *Ear Hear*, vol. 31, no. 3, pp. 345–355, 2010.

[43] J. A. Maxwell and P. M. Zurek, “Reducing Acoustic Feedback in Hearing Aids,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 304–313, jul 1995.

[44] A. Spriet, M. Moonen, and J. Wouters, “Evaluation of Feedback Reduction Techniques in Hearing Aids Based on Physical Performance Measures,” *The Journal of the Acoustical Society of America*, vol. 128, no. 3, p. 1245, 2010.

[45] C. Lee, J. M. Kates, B. D. Rao, and H. Garudadri, “Speech Quality and Stable Gain Trade-Offs in Adaptive Feedback Cancellation for Hearing Aids,” *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. EL388–EL394, 2017.

[46] “Hearing Assistive Technology.” [Online]. Available: http://www.asha.org/public/hearing/hearing-assistive-technology.

[47] W. M. Whitmer, C. G. Brennan-Jones, and M. A. Akeroyd, “The Speech Intelligibility Benefit of a Unilateral Wireless System for Hearing-Impaired Adults.” *International journal of audiology*, vol. 50, no. 12, pp. 905–11, Dec. 2011.

[48] G. M. Craddock, L. P. McCormack, R. B. Reilly, and H. T. Knops, *Assistive Technology - Shaping the Future*. IOS Press, 2003.

[49] E. C. Schafer, K. Sanders, D. Bryant, K. Keeney, and N. Baldus, “Effects of Voice Priority in FM Systems for Children with Hearing Aids,” *Journal of Educational Audiology*, vol. 19, pp. 12–24, 2013.

[50] L. Thibodeau, “Comparison of Speech Recognition with Adaptive Digital and FM Remote Microphone Hearing Assistance Technology by Listeners Who Use Hearing Aids,” *American Journal of Audiology*, vol. 23, pp. 201–210, June 2014.

[51] W. Lin, D. Tao, J. Kacprzyk, Z. Li, E. Izquierdo, and H. Wang, Eds., *Multimedia Analysis, Processing and Communications*. Springer, 2011.

[52] 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, 2006.

[53] N. Côté, "Speech Quality Measurement Methods," in *Integral and Diagnostic Intrusive Prediction of Speech Quality*. Springer Berlin Heidelberg, 2011, no. 1997, ch. 2.

[54] N. Pourmand, V. Parsa, and A. Weaver, "Computational Auditory Models in Predicting Noise Reduction Performance for Wideband Telephony Applications," *International Journal of Speech Technology*, vol. 16, no. 4, pp. 363–379, 2013.

[55] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE transactions on audio, speech and language processing*, vol. 16, no. 1, pp. 229–238, 2008.

[56] D. G. Altman and J. M. Bland, "Measurement in Medicine: the Analysis of Method Comparison Studies," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 32, no. 3, pp. 307–317, 1983.

[57] J. H. Steiger, "Tests for Comparing Elements of a Correlation Matrix." pp. 245–251, 1980.

[58] Hu, Yi and Loizou, Philipos C., "Subjective Comparison of Speech Enhancement Algorithms," *in Proc. IEEE ICASSP*, pp. 153–156, 2006.

[59] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Quality Index (HASQI) Version 2," *AES: Journal of the Audio Engineering Society*, vol. 62, no. 3, pp. 99–117, 2014.

[60] D. Suelzle, V. Parsa, and T. H. Falk, "On a Reference-Free Speech Quality Estimator for Hearing Aids." *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 412–418, 2013.

[61] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Objective Perceptual Quality Measures for the Evaluation of Noise Reduction Schemes," *9th International Workshop on Acoustic Echo and Noise Control*, pp. 169–172, 2005.

[62] A. A. Kressner, D. V. Anderson, and C. J. Rozell, "Robustness of the Hearing Aid Speech Quality Index (HASQI)," *IEEEWorkshop on Applications of Signal Processing to Audio and Acoustics*, pp. 209–212, 2011.

[63] P. C. Loizou, *Speech Enhancement: Theory and Practice, Second Edition*. CRC Press, 2007.

[64] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Prentice Hall, 1988.

[65] G. Chen, "Statistical Model-Based Objective Measures of Speech Quality Certificate of Examination," Ph.D. dissertation, University of Western Ontario, 2007.

[66] N. Pourmand, "Objective and Subjective Evaluation of Wideband Speech Quality," Ph.D. dissertation, University of Western Ontario, 2012.

[67] S. Wang, A. Sekey, and A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 819–829, 1992.

[68] B. Boehm, C. Abts, and S. Chulani, "Software Development Cost Estimation Approaches-A Survey," *Annals of software engineering*, vol. 10, pp. 177–205, 2000.

[69] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, P. A. Naylor, M. Brookes, and M. Huckvale, "Data Driven Method for Non-Intrusive Speech Intelligibility Estimation," *European Signal Processing Conference*, pp. 1899–1903, 2010.

[70] G. Chen and V. Parsa, "Bayesian Model Based Non-Intrusive Speech Quality Evaluation," *Proc. International Conference Acoustics, Speech, and Signal Processing*, pp. 385–388, 2005.

[71] N. Pourmand, D. Suelzle, V. Parsa, Y. Hu, and P. Loizou, "On the Use of Bayesian Modeling for Predicting Noise Reduction Performance," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3873–3876.

[72] T. H. Falk, Q. Xu, and W.-Y. Chan, "Non-Intrusive GMM-Based Speech Quality Measurement," *International Conference on Acoustics, Speech, and Signal Processing,ICASSP*, pp. 125–128, 2005.

[73] T. H. Falk and W.-Y. Chan, "Single-Ended Speech Quality Measurement Using Machine Learning Methods," *IEEE Transactions On Audio Speech And Language Processing*, vol. 14, no. 6, pp. 1935–1947, 2006.

[74] J. H. Friedman, "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–67, 1991.

[75] H. Salehi and V. Parsa, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

[76] B. Scholkopf and A. Smola, *Learning With Kernels*. Cambridge, MA: MIT Press, 2002.

[77] M. Davy and S. Godsill, "Detection of Abrupt Spectral Changes Using Support Vector Machines an Application to Audio Signal Segmentation," *in Proc. IEEE International Conference Acousticts, Speech, Signal Processing*, no. 2, pp. 1313–1316, 2002.

[78] S. Chen, R. Guido, T. Truong, and Y. Chang, "Improved Voice Activity Detection Algorithm Using Wavelet and Support Vector Machine," *Computer Speech Language*, vol. 24, no. 3, pp. 531–543, 2010.

[79] W. Campbell, J. Campbell, T. P. Gleason, D. A. Reynolds, and W. Shen, "Speaker Verification Using Support Vector Machines and High-Level Features," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2085–2094, 2007.

[80] K. Aida-zade, A. Xocayev, and S. Rustamov, "Speech Recognition Using Support Vector Machines," *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, vol. 1, pp. 1–4, 2016.

[81] M. Narwaria, W. Lin, S. Member, I. V. Mcloughlin, S. Member, S. Emmanuel, and L.-T. Chia, "Nonintrusive Quality Assessment of Noise Suppressed Speech With Mel-Filtered Energies and Support Vector Regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1217–1232, 2012.

[82] A. K. Jain, R. P. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[83] J. Shlens, "A Tutorial on Principal Component Analysis," 2014.

[84] R. E. Crochiere, J. M. Tribolet, and L. R. Rabiner, "An Interpretation of the Log-Likelihood Ratio as a Measure of Waveform Coder Performance," *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. ASSP-28, no. 3, pp. 318–323, 1980.

[85] L. Gu, J. G. Harris, R. Shrivastav, and C. Sapienza, "Disordered Speech Evaluation Using Objective Quality Measures," *ICASSP*, pp. 321–324, 2005.

[86] N. Jayant and P. Noll, *Digital Coding of Waveforms:Principles and Applications to Speech and Video*. Prentice-Hall, 1984.

[87] H. P. Knagenhjelm and W. B. Kleijn, "Spectral Dynamics Is More Important Than Spectral Distortion," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 732–735, 1995.

[88] J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of Spectral Centroid Magnitude and Frequency for Speaker Recognition," *The Speaker and Language Recognition Workshop*, pp. 34–39, 2010.

[89] D. Sharma and P. A. Naylor, "Evaluation of Pitch Estimation in Noisy Speech for Application in Nonintrusive Speech Quality Assessment," *Proc. European Signal processing Conf*, pp. 2514–2518, 2009.

[90] D. Sharma, "Speech Assessment and Characterization for Law Enforcement Applications," Ph.D. dissertation, Imperial College London, 2012.

[91] H. J. McDermott, "A Technical Comparison of Digital Frequency-Lowering Algorithms Available in Two Current Hearing Aids," *PLoS ONE*, vol. 6, no. 7, 2011.

[92] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes, “A Data-Driven Non-Intrusive Measure of Speech Quality and Intelligibility,” *Speech Communication*, vol. 80, pp. 84–94, 2016.

[93] K. Kondo, “Speech Quality,” in *Subjective Quality Measurement of Speech, Its Evaluation, Estimation and applications*, ser. Signals and Communication Technology. Springer Berlin Heidelberg, 2012, ch. 2, pp. 8–10.

[94] ITU-T, “Rec. P.863, Perceptual Objective Listening Quality Assessment,” 2011.

[95] R. Huber and B. Kollmeier, “PEMO-Q A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[96] R. Huber, V. Parsa, and S. Scollie, “Predicting the Perceived Sound Quality of Frequency-Compressed Speech,” *PLoS ONE*, vol. 9, no. 11, p. e110260, 2014.

[97] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, “ViSQOL: an Objective Speech Quality Model,” *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 13, no. 1, 2015.

[98] M. S. A. Zilany, I. C. Bruce, P. C. Nelson, and L. H. Carney, “A Phenomenological Model of the Synapse Between the Inner Hair Cell and Auditory Nerve: Long-Term Adaptation with Power-Law Dynamics,” *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2390–2412, 2009.

[99] M. R. Wirtzfeld, N. Pourmand, V. Parsa, and I. C. Bruce, “Predicting the Quality of Enhanced Wideband Speech with a Cochlear Model,” *The Journal of the Acoustical Society of America 142,*, vol. 319, no. 905, pp. 1–22, 2017.

[100] J. M. Kates and K. H. Arehart, “The Hearing-Aid Speech Quality Index (HASQI),” *J. Audio Eng. Soc.*, vol. 58, no. 5, pp. 363–381, 2010.

[101] M. S. A. Zilany, I. C. Bruce, and L. H. Carney, “Updated Parameters and Expanded Simulation Options for a Model of the Auditory Periphery,” *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 283–286, 2014.

[102] F. Chen, O. Hazrati, and P. C. Loizou, “Predicting the Intelligibility of Reverberant Speech for Cochlear Implant Listeners with a Non-Intrusive Intelligibility Measure.” *Biomedical signal processing and control*, vol. 8, no. 3, pp. 311–314, 2013.

[103] T. H. Falk, C. Zheng, and W.-Y. Chan, “A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[104] J. Kates, “An Auditory Model for Intelligibility and Quality Predictions,” *Proceedings of Meetings on Acoustics*, vol. 19, pp. 1–9, 2013.

[105] J. Makhoul, S. Roucos, and H. Gish, "Vector Quantization in Speech Coding," *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1551–1588, 1985.

[106] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech." *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–52, 1990.

[107] H. Salehi and V. Parsa, "Nonintrusive Speech Quality Estimation Based on Perceptual Linear Prediction," *IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pp. 1–4, 2016.

[108] M. Slaney, "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank," Tech. Rep. 35, 1993. [Online]. Available: https://engineering.purdue.edu/malcolm/apple/tr35/PattersonsEar.pdf

[109] P. Kabal, "TSP Speech Database," 2002.

[110] G. Stoll and F. Kozamernik, "EBU Listening Tests on Internet Audio Codecs," *EBU Technical Review*, no. June, p. 24, 2000.

[111] G. Jekabsons, "ARESLab: Adaptive Regression Splines Toolbox for Matlab/Octave," pp. 1–19, 2016. [Online]. Available: http://www.cs.rtu.lv/jekabsons/

[112] D. Suelzle, "Electroacoustic and Behavioural Evaluation of Hearing Aid Digital Signal Processing Features," Ph.D. dissertation, University of Western Ontario, 2013.

[113] K. Smeds, F. Wolters, and M. Rung, "Estimation of Signal-to-Noise Ratios in Realistic Sound Scenarios," *Journal of the American Academy of Audiology*, vol. 26, no. 2, pp. 183–196, 2015.

[114] N. Bisgaard, M. Vlaming, and M. Dahlquist, "Standard Audiograms for the IEC 60118-15 Measurement Procedure," *Trends in Amplification*, vol. 14, no. 2, pp. 113–120, 2010.

[115] I. Holube, S. Fredelake, M. Vlaming, and B. Kollmeier, "Development and Analysis of an International Speech Test Signal (ISTS)," *International Journal of Audiology*, vol. 49, no. 12, pp. 891–903, 2010.

[116] V. Parsa, S. Scollie, D. Glista, and A. Seelisch, "Nonlinear Frequency Compression: Effects on Sound Quality Ratings of Speech and Music," *Trends in Amplification*, vol. 17, no. 1, pp. 54–68, 2013.

[117] S. Scollie, R. Seewald, L. Cornelisse, S. Moodie, M. Bagatto, D. Laurnagaray, S. Beaulac, and J. Pumford, "The Desired Sensation Level Multistage Input/Output Algorithm," *Trends in Amplification*, vol. 9, no. 4, pp. 159–197, 2005.

[118] "Comfort Audio AB. Digital Communication Systems for Hearing Implants." [Online]. Available: http://www.comfortaudio.com/wp-content/blogs.dir/9/files_mf/kompabilitetsguide_ci_eng_120314.pdf

[119] "Oticon Pediatrics. Product Information - Amigo T30/T31." [Online]. Available: http://www.oticon.com.br/~asset/cache.ashx?id=10589&type=14&format=web

[120] "Phonak Easylink Product Information." [Online]. Available: https://www.phonakpro.com/content/dam/phonakpro/gc_hq/en/products_solutions/wireless_accessories/phase_out/com_datasheet_easylink.pdf

[121] "Roger Inspiro Overview | PhonakPro." [Online]. Available: https://www.phonakpro.com/ca/en/products/wireless-accessories/roger-inspiro/overview-roger-inspiro.html

[122] J. Wolfe, M. M. Duke, E. Schafer, C. Jones, H. E. Mülder, A. John, and M. Hudsone, "Evaluation of Performance with an Adaptive Digital Remote Microphone System and a Digital Remote Microphone Audio-Streaming Accessory System," *Journal of Speech, Language, and Hearing Research*, vol. 24, no. 2, pp. 1–14, 2015.

[123] K. H. Arehart, J. M. Kates, and M. C. Anderson, "Effects of Noise, Nonlinear Processing, and Linear Filtering on Perceived Music Quality," *International Journal of Audiology*, vol. 50, no. 3, pp. 177–190, 2011.

[124] J. M. Kates, S. Member, and K. H. Arehart, "The Hearing-Aid Audio Quality Index (HAAQI)," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 24, no. 2, pp. 354–365, 2016.

CURRICULUM VITA
HANIYEH SALEHI

EDUCATION

2012-2017	**The University of Western Ontario**
Degree: **Ph.D. in Electrical Engineering**	**London, Canada**
2008-2011	**K.N. Toosi University of Technology**
Degree: **M.Sc. in Electrical Engineering**	**Tehran, Iran**
2002-2007	**The University of Shahid Beheshti**
Degree: **B.Sc. in Electrical Engineering**	**Tehran, Iran**

THESES

- Ph.D. Thesis
  Learning-Based Reference-Free Speech Quality Assessment for Normal Hearing and Hearing Impaired Applications
  Advisor: Dr. Vijay Parsa
- M.Sc. Thesis
  Design of an Application-Specific Controller for a Visual Prosthesis
  Advisor: Dr. Amir Masoud Sodagar
- B.Sc. Thesis
  Design and Prototype of a Resonant Converter
  Advisor: Dr. Ebrahim Afjei

TEACHING EXPERIENCE

2013-2015	**The University of Western Ontario**
Position: **Teaching Assistant**	**London, Canada**

WORK EXPERIENCE

2012–2017	**National Centre for Audiology**
Position: **Research Assistant**	**London, Canada**
2016–2017	**ON Semiconductor**
Position: **Evaluation Engineer**	**Waterloo, Canada**
2008–2011	**Integrated Circuits and Systems (ICAS) Lab.**
Position: **Research Assistant**	**Tehran, Iran**

HONORS AND AWARDS

- Western Graduate Research Scholarship (WGRS), UWO, London, Canada [2012-2016].

- Electrical and Computer Engineering Graduate Student Travel Award, UWO [2016].

JOURNAL PAPERS

- **H. Salehi**, V. Parsa, and P. Folkeard “Electroacoustic Assessment of Wireless Remote Microphone Systems”, *Audiology Research*, vol. 8, no. 1, Apr. 2018.
- **H. Salehi**, D. Suelzle, P. Folkeard, and V. Parsa, “Learning-Based Reference-Free Speech Quality Measures for Hearing Aid Applications”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (Under Revision).

REFEREED CONFERENCE PAPERS

- **H. Salehi** and V. Parsa, “Nonintrusive Speech Quality Estimation Based on Perceptual Linear Prediction”, *IEEE Canadian Conference on Electrical and Computer Engineering (CCECE16), Vancouver, Canada, May. 2016.*
- **H. Salehi** and V. Parsa, “On Nonintrusive Speech Quality Estimation for Hearing Aids”, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, U.S., Oct. 2015.*
- **H. Salehi** and A.M. Sodagar, “Muti-Mode Application-Specific Controller Dedicated to a Visual Prosthesis”, *8th International Caribbean Conference on Devices, Circuits and Systems (ICCDCS), Playa del Carmen, Mexico, Mar. 2012.*

PEER REVIEWED ABSTRACTS

- **H. Salehi**, V. Parsa, P. Folkeard, S. Steffler, and J. Ryan, “Reference-Free Metrics for Hearing Aid Speech Quality Assessment”, *International Hearing Aid Research Conference, Aug. 2016.*
- D. Glista, M. Hawkins, **H. Salehi**, N. Pourmand, V. Parsa, and S. Scollie, “Evaluation of Sound Quality with Adaptive Nonlinear Frequency Compression”, *International Hearing Aid Research Conference, Aug. 2016.*
- **H. Salehi** and V. Parsa, “Perceptual Linear Prediction Incorporating Hearing Loss Model”, *Speech and Audio in the Northeast (SANE), Google, New York City, U.S., Oct. 2015.*
- **H. Salehi**, V. Parsa, and Paula Folkeard, “On Electroacoustic Assessment of Remote Microphone Devices/Systems”, *International Hearing Aid Research Conference”, Aug. 2014.*
- **H. Salehi** and A.M. Sodagar, “An ASIC Controller for a Visual Prosthesis”, *Farabi Festival for Ophthalmology and Visual Sciences, Tehran, Iran, Mar. 2012.*