Addis Ababa University
Addis Ababa Institute of Technology
School of Electrical and Computer Engineering

Perceptual Objective Audio Quality Assessment using Computational Auditory Model (POAQ-CAM)

**A thesis submitted to Addis Ababa Institute of Technology,
School of Graduate Studies, Addis Ababa University
In Partial Fulfillment of the Requirements for the Degree of Masters of Science in
Electrical Communication Engineering**

By

Rediet Million

Advisor: Dr.-Ing. Dereje Hailemariam

November 2019

Addis Ababa, Ethiopia

Addis Ababa University
Addis Ababa Institute of Technology
School of Electrical and Computer Engineering

Perceptual Objective Audio Quality Assessment using Computational Auditory Model (POAQ-CAM)

By

Rediet Million

Approval by Board of Examiners

Dr.Yalemzewed Negash Dean, School of Electrical and Computer Engineering	Signature
Dr.Ephrem Teshale Chairman	Signature
Dr.-ing. Dereje Hailemariam Advisor	Signature
Dr.Enyew Adugna Internal Examiner	Signature
Dr. Dawit Assefa External Examiner	Signature

Declaration

I, the undersigned, declare that this thesis work, to the best of my knowledge and belief, is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been fully acknowledged.

Rediet Million _____

Name Signature

Place: Addis Ababa, Ethiopia

Date of Submission: November 1, 2019

This thesis has been submitted for examination with my approval as a university advisor.

Dr.-Ing. Dereje Hailemariam _____

Advisor's name Signature

Acknowledgment

First and foremost, I would like to thank the Almighty God for giving me the strength and making me courageous during my study. Secondly, my deepest gratitude goes to my thesis advisor, Dr.-Ing. Dereje Hailemariam, for his valuable comments, suggestion, and guidance which helped me a lot in pursuing my thesis work.

I would like to appreciate my colleagues at the School of Electrical and Computer Engineering, and friends for their support and suggestion in the course of my MSc study and thesis work. I would like to thank Prof.Andrew Hines from Dublin Institute of Technology, Ireland, for sharing me valuable resources that helped me a lot in my thesis work.

Finally, I want to express my grateful appreciation to my mother, W/ro Addis W/T-sadik, for her love, support, encouragement, and everything else during those times when I needed it the most. There are so many others whom I may have unintentionally left out and I sincerely thank all of them for their help.

Abstract

Audio quality assessment is one of the key considerations in television and radio broadcasting systems. Up until now, in many Ethiopian broadcasting service providers, the only way to measure the audio quality of an end to end broadcasting system is by using the human subjective evaluation method. Even though subjective evaluation is the ideal method for assessing the audio quality, it is time-consuming and even more challenging for real-time audio quality monitoring. We, therefore, require an objective measurement method that can quantitatively estimate the perceptual audio quality based on the physical features of the audio signal. Consequently, the idea of substituting the subjective evaluation by objective, computer-based methods has been an ongoing focus of research and development.

Several methods for making objective perceptual measurement of audio quality have been introduced and standardized during the last decade; however, most approaches have accuracy issues. To solve these problems an improved intrusive objective audio quality estimation system has been proposed based on a psychoacoustically validated computational human auditory model and based on the framework of the VISQOLAudio audio quality assessment metric.

This thesis evaluates the performance of the proposed model, POAQ-CAM, against the latest ViSQOLAudio, PEAQb and PEMO-Q models using a large database of audio and speech subjective listening tests that were originally carried out by International Telecommunication Union (ITU), CoreSV and TCD-VOIP. Compared to the above models, the proposed estimation system has a considerable improvement both in terms of accuracy, measured using root mean square error (RMSE) and linearity measured based on the correlation coefficient. The RMSE values of POAQ-CAM are reduced by 22.65%, 43.1% and 48.1% with respect to ViSQOLAudio, PEMO-Q and PEAQb models, while the correlation is increased by 6% and 3.9% compared to PEAQb and PEMO-Q models.

Key words: POAQ-CAM, Audio quality assessment, Computational auditory model

Contents

Acknowledgment iv

Abstract v

List of Figures **viii**

List of Tables **ix**

Acronyms **x**

1 Introduction **1**

1.1 Statement of the Problem 2

1.2 Objectives 3

1.2.1 General Objective 3

1.2.2 Specific Objectives 3

1.3 Significance of the Study 3

1.4 Scope 4

1.5 Thesis Outline 4

2 Literature Review **5**

2.1 Review on Objective Audio Quality Metrics 5

2.1.1 PEAQ 6

2.1.2 PEMO-Q 7

2.1.3 ViSQOLaudio 8

3 Fundamental Concepts **10**

3.1 Approaches to Audio Quality Assessment 10

3.1.1 Subjective Assessment 11

3.1.2 Objective Assessment 13

3.2 The Human Auditory System 14

3.2.1 Outer and Middle Ear 15

3.2.2 Inner Ear 16

4 Methodology **19**

4.1 Methodology for Designing the New System 19

4.1.1 Pre-Processing 19

4.1.2 Auditory Processing 22

4.1.3 Post-Processing: Modeling the Cognitive Effects 27

4.2 Experiment Methodology 30

4.2.1 Experimental Datasets 31

4.2.2 Objective Models Configuration 32

4.2.3 Performance Evaluation Metrics 33

5 Results and Discussion **35**

5.1 Simulation Results and Discussion 35

5.2 Performance Evaluation 40

6 Conclusion and Recommendations **44**

6.1 Conclusion 44

6.2 Recommendations 45

Bibliography **46**

List of Figures

2.1 A high-level representation of ViSQOLAudio [34]. 9

3.1 Types of objective audio quality assessment. 13
3.2 Anatomy of the human ear [28]. 15
3.3 Cross section of the cochlear [28]. 16
3.4 Resonance structure of basilar membrane in unrolled cochlear [28]. 17
3.5 Organ of corti, inner hair cell and its equivalent circuit [28]. 18

4.1 System architecture of the proposed POAQ-CAM system. 20
4.2 Block diagram of the auditory processing adopted from CASP model. 23
4.3 The process of creating an NSIM patch by comparing the similarity of a reference and degraded patch pair [34]. 28
4.4 The process of generating a MOS-LQO from NSIM patches [34]. 30

5.1 Stereo audio input (left panel) and output (right panel) of mid channel extraction process. 35
5.2 Outer and middle ear frequency response. 36
5.3 Input signal from pre-processing stage (left)and the corresponding output of the outer-middle ear processor(right). 37
5.4 Gammatone filterbank processor. 38
5.5 Comparison between the outputs of gammatone and DRNL processor. 39
5.6 Illustration of the envelope extraction processor. BM output (left) and the corresponding IHC model output (right). 40
5.7 Illustration of the adaptation processor. IHC output (left) as the input to the adaptation processor and the corresponding output using(right). 40
5.8 Histogram of RMSE and R-correlation score. 43
5.9 Subjective versus objective quality scores. First and third order polynomial regression curves. 43

List of Tables

3.1 ACR test grade in MOS scale. 11

3.2 DCR test grade in DMOS scale. 12

3.3 CCR test grade in CMOS scale. 12

5.1 Results of SDG and ODG scores. 42

5.2 Result of RMSE and correlation score. 42

Acronyms

CMOS: Comparison Mean Opinion Score

ACR: Absolute Category Rating

AN: Auditory Nerve

ASD: Auditory Spectral Difference

BM: Basilar Membrane

BSD: Bark Spectral Distance

CASP: Computational Auditory Signal Processing

CF: Characteristic Frequencies

DCR: Degradation Category Rating

DFT: Discrete Fourier Transform

DIX: Disturbance Index

DMOS: Degradation Mean Opinion Score

DRNL: Dual Resonance Non-Linear

EBC: Ethiopia Broadcast Corporation

ERB: Equivalent Rectangular Bandwidth

FFT: Fast Fourier Transform

IHC: Inner Hair Cell

ITU: International Telecommunications Union

MOS: Mean Opinion Score

MOS-LQO: Mean Opinion Score-Listening Quality Objective

MOS-LQS: Mean Opinion Score-Listening Quality Subjective

MUSHRA: MUlti Stimulus test with Hidden Reference and Anchor

NMR: Noise to Mask Ratio

NSIM: Neurogram Similarity Index Measure

ODG: Objective Difference Grade

OPS: Overall Perception Score

PAQM: Perceptual Audio Quality Measure

PEAQ: Perceptual Evaluation of Audio Quality

PEMO-Q: PEreceptual Model-Quality

PERCEVAL: PERCeptual EVALuation

POAQ-CAM: Perceptual Objective Audio Quality-Computational Auditory Model

POLQA: Perceptual Objective Listening Quality Analysis

POM: Perceptual Objective Measure

PSQM: Perceptual Speech Quality Measure

QoE: Quality of Experience

QoS: Quality of Service

RMSE: Root Mean Square Error

SDG: Subjective Difference Grade

SPL: Sound Pressure Level

SVR: Support Vector Regression

ViSQOL: Virtual Speech Quality Objective Listener

Chapter 1

Introduction

Ethiopia is presently undergoing dramatic changes in the expansion of broadcast industries both in private and public sectors. Currently, ten radio stations, twenty-four satellite, and terrestrial television channels are licensed to operate in the country[1]. The production and transmission techniques of these service providers are based on new digital broadcasting system technologies. New technologies bring new challenges with it, and the move to a digital-based system means that engineers have to acquire new, reliable means of testing their system quality.

Broadcasting system is a long and complex path involving electro-acoustic transducers, mixers, storage media, interface channels, signal processing and so on. Equipment used within this chain will be the limiting factor of performance. Dynamic compression will generate harmonic distortion; noise from microphone preamps may limit system dynamic range; lossy compression codecs will exhibit complex noise and distortion characteristics [18] which, as a result, have a significant impact on the end-user perceived audio quality.

End-to-end evaluation of the audio quality has become more complex as the number of variables impacting the signal has expanded. Subjective evaluation with human listeners is a ground truth measurement for audio quality but it is time-consuming and expensive to carry out [2]. Also, audio analyzer tools can give a good indicator of the quality of service (QoS), but predicting the quality of experience (QoE) is becoming more important for the end-user. Thus, computer-based objective measures aim to model this assessment to give accurate estimates of quality when compared with subjective evaluations.

Many objective algorithms have been proposed in the literature and some of them have been standardized. Since the advent of the modern perceptual audio codec in the late

1980s, many objective quality metrics have been developed that try to measure human subjective audio quality [38][3][25][32]. The International Telecommunications Union (ITU) has put forth recommendation BS.1387, also known as Perceptual Evaluation of Audio Quality (PEAQ) as an approach for doing exactly this [32]. PEAQ combines many of the best available quality metrics available in the early 1990s, attempting to merge their various strengths. PEAQ, however, has been shown to be a poor indicator of perceptual quality for highly impaired audio [9]. More recently, a new method called Virtual Speech Quality Objective Listener (ViSQOL) [16], which is later adapted to a function as a perceptual audio quality model ViSQOLAudio[34], has been introduced. It is a signal-based, full reference, an intrusive metric that model human audio quality perception using Neurogram Similarity Index Measure (NSIM) between a reference and test audio signals [16]. Compared to the existing models, ViSQOLAudio has been shown to be especially effective as a measure of the quality of highly impaired audio. However, this model doesn't exploit the very fundamental properties of the human auditory system that most advanced perceptual objective audio quality assessment model does.

Thus, the goal of this thesis is to implement and verify an improved objective audio quality assessment method based on a psychoacoustically validated computational human auditory model introduced by *M. Jepsen et al.*[19] and based on the framework of ViSQOLAudio model introduce by *A.Hines et al.* [34].

1.1 Statement of the Problem

Ethiopia Broadcast Corporation (EBC) is one of the biggest broadcast service provider in Ethiopia. It has three public TV channels and one FM radio station. Most of them are playing out for 24 hours a day and almost half of the programs are newly produced. That means when the system is running at its full capacity around 100 hours a new program to be broadcasted every day. Engineers monitor the system in real-time and verify the audio signal using a set of the audio analyzers. It is labor-intensive, one engineer could only inspect one program at one time and quality issues may be omitted by mistakes.

Many literatures show that predicting perceived audio quality degradation turned out to be more difficult for general wide-band audio signals. Although some objective audio quality assessment methods perform well, none of them fulfilled the requirements of accuracy level. Recently, Dr. Andrew and his colleagues have done a well-performing model than all other existing models [34]. However, this model doesn't exploit the very fundamental properties of the human auditory system which has a direct effect on the results of the perceptual audio quality assessment score. Therefore, the goal of this thesis is to design and implement an improved objective audio quality assessment method based on a human auditory model that can increase the level of accuracy.

1.2 Objectives

1.2.1 General Objective

The main objective of this thesis research project is to design and implement an improved objective audio quality assessment model that can estimate the audio quality of the broadcasting system.

1.2.2 Specific Objectives

- To study and understand psychoacoustic principles of the human auditory system.
- To develop an improved algorithm by adopting a computational human auditory model into an existing framework of ViSQOLAudio metrics.
- To test and evaluate the new approach using different audio and speech datasets and compare the results with existing benchmark models.
- To draw a conclusions based on the result obtained from performance evaluation metrics.

1.3 Significance of the Study

This study has the following benefits:

- To develop local-based audio quality assessment tools and facilitate technology-transfer for digital broadcasting, telecommunication and multimedia sectors.
- To decrease the cost and time of a subjective based quality assessment process.

1.4 Scope

The main focus of this study is to design and implement perpetually improved objective audio quality metrics based on the psychoacoustically validated model of auditory processing. Even though there are a few numbers of non-reference objective audio quality evaluation metrics, the current study is limited on full-reference based audio quality metrics which measure the audio quality difference between reference and test signals.

1.5 Thesis Outline

The rest of the thesis has been organized into five Chapters. *Chapter 2* discusses the work of related researches in the area of full reference objective audio quality assessment. *Chapter 3* provides fundamental concepts related to the audio quality assessment approaches and the human auditory system. The methodology for designing the proposed model is discussed in *Chapter 4*. The result and discussion of the evaluation are presented in *Chapter 5*. Finally, conclusions and recommendations are provided in *Chapter 6*.

Chapter 2

Literature Review

This chapter presents the relevant literatures on audio quality assessment models from the past to recent development. Metrics such as PEAQ [32], PEMO-Q [17] and ViSQOLAudio[34] will be investigated. These metrics are given particular attention because they will be part of the performance evaluation experiments in Chapter five.

2.1 Review on Objective Audio Quality Metrics

An objective audio quality assessment has been of great interest to researchers for many years since the end of the 1940s. In the year 1979, Schroeder *et al.* [33] proposed the concept of noise loudness and unprecedentedly used a simple masking method to estimate the audibility of coding noise in a speech coder, starting to apply the auditory perception process to the sound quality evaluation. Compared with Schroeder's approach, Karjalainen (1985) [20] improved the accuracy of assessment by ASD (Auditory Spectral Difference) model that can be adapted to simulate a much wider range of perceptual effects, thereby his approach has been much more successful and dominant in this field.

In 1987, Brandenburg [5] published the measurement scheme NMR (Noise to Mask Ratio), which added a simple scheme of simulating post-masking and several ways to evaluate the perceived quality of audio encoding at low bit rates. NMR makes a measure of the level difference between the masked threshold and the noise signal. The frequency content of the signal is analyzed by taking a Discrete Fourier Transform (DFT) with a Hann window of 20 ms duration and the transform coefficients are mapped to a bark scale. The masked threshold is estimated for each band and the slope of the masked threshold is obtained using a worst-case approach. NMR was the first one implemented in real-time hardware among the audio quality methods [32].

In the early 1990s, a similar approach to the ASD called Bark Spectral Distance (BSD)[40] was introduced by Wang *et al.* BSD computes the mean squared Euclidean distance on a sone scale in the bark bands but it does not take temporal masking into account. After that, Beerends and Stemerdink's (1992) [3] perceptual audio quality measure (PAQM) was introduced. The basic idea of PAQM is to subtract the 'internal representations' of the reference and degraded signal and map the difference with a cognitive mapping to the subjectively perceived audio quality. To achieve greater accuracy, PAQM was integrated with NMR and then it was adapted into a method for speech coder evaluation known as PSQM (Perceptual Speech Quality Measure), which was later adopted as PESQ. Then, in 1995, Paillard *et al.* [8] published the PERCeptual EVALuation (PERCEVAL), which models the transfer characteristics of the middle and inner ear to form an internal representation of the signal.

2.1.1 PEAQ

In 1994, the International Telecommunications Union (ITU) created a task group whose goal was to develop a recommendation for objectively measuring perceived audio quality. An open call of proposals was issued and six perceptual measurement methods were presented: Disturbance Index (DIX) [38], Noise-to-Mask Ratio (NMR) [5], Perceptual Audio Quality Measure (PAQM)[3], Perceptual Evaluation (PERCEVAL) [25], Perceptual Objective Measure (POM) [8], and the Toolbox Approach [39][32]. The best features of the original methods were combined into a new objective model for evaluating sound quality called Perceptual Evaluation of Audio Quality (PEAQ), and formally standardized as ITU-R Recommendation BS.1387 [32].

There are two versions of PEAQ: a basic version, which is optimized for speed, and an advanced model that adds a filterbank based ear model to the basic FFT-based model to improve accuracy. The PEAQ quality prediction process begins by passing the reference and degraded signals into an ear model that segments the signals into auditory filter

bands that, among other steps are passed into weighted transfer functions representing the different parts of the ear. A process then identifies excitation patterns in loudness and modulation. These patterns are used to calculate several psycho-acoustically based model output variables, such as average linear distortions, that quantify differences between the reference and degraded signals. These model output variables and a set of coefficients are inputted to an artificial neural network which outputs a distortion index that is mapped to an Objective Difference Grade (ODG). An ODG is analogous to the Subjective Difference Grade, where the grade is the ITU-T BS.562. impairment scale from 1 (very annoying) to 5 (imperceptible)[34].

A decade later, ITU-T Recommendation P.863 standardized a new objective metric for measuring speech quality, called POLQA. POLQA was designed for speech quality assessment and can be run in narrowband mode (telephone quality; 300–3400 Hz) or super wideband (SWB) mode (50–14 000 Hz). POLQA has been shown to have potential as an audio quality model and the developers of POLQA are currently working on adapting it for use in audio quality evaluation [34].

2.1.2 PEMO-Q

Another model for predicting perceptual audio quality, PEMO-Q [17], was shown by its authors to predict quality more accurately than PEAQ. PEMO-Q predicts quality using time-aligned reference and degraded signals that are level aligned before deleting silence from the signals. The signals are input to a psychoacoustically motivated model that transforms the signals into a three-dimensional representation, where the dimensions represent activity patterns in time, frequency and modulation-frequency. The correlations between the reference and degraded patterns are used to create error estimations that are divided into target distortion, interference and artifact components. Each component is weighted for salience and the weights are input to a trained non-linear mapping that produces an Overall Perception Score(OPS) value ranging from 0 (bad quality) to 100 (excellent quality).

2.1.3 ViSQOLaudio

Recently, at QxLab [27], by Andrew H. et al.[34], an alternative speech and audio quality model called ViSQL(Virtual Speech Quality Objective Listener) has been developed for quality assessment in narrow to full band mode. It is the first free and open-source audio quality metric with accuracy comparable to models used in the industry.

The quality score prediction process of ViSQOLAudio has four phases: *pre-processing, pairing, comparison, and similarity to quality mapping.* A high-level representation of the model is given in Figure 2.1. In the *pre-processing* phases, the mid-channel is extracted from the reference and degraded signal to consider information from both channels. An alignment process is then performed on the reference and degraded signals, compensating for subframe misalignment caused by encoder padding. A spectrogram of the reference and degraded signals is then built using a Gammatone filter. The *pairing phase* first segments the reference spectrogram into patches of 30 frames. These patches are used as input into a robust alignment process that matches each reference spectrogram patch with the most similar patch from the degraded spectrogram, creating a set of most similar reference-degraded patch pairs. This alignment process helps to correct drift and warping in the degraded signal. In the *comparison phase*, the similarity of each most similar patch pair is measured, outputting similarity patches representing the similarity of each of the pairs. For each of these similarity patches, the similarity across each a frequency band is measured. This allows each of the similarity of each frequency band in the degraded signal to be considered separately, allowing a machine learning model to find relationships between similarities across frequency bands that are used to make more accurate quality score predictions. The similarity to *quality mapping phase* inputs the mean frequency band similarity scores of each similarity patch into a Support Vector Regression (SVR) model that outputs a MOS-LQO value[34].

Compared to the previous models, ViSQOLAudio is especially effective as a measure of the quality of highly impaired audio. However, this models doesn't exploit the very fun-

damental properties of the human auditory system that the most advanced perceptual objective audio quality assessment models do. In this thesis, by adopting the latest Computational Auditory Signal Processing (CASP) algorithm into the framework of ViSQOLAudio model, an improved objective audio quality assessment metrics, POQA-CAM, is introduced.

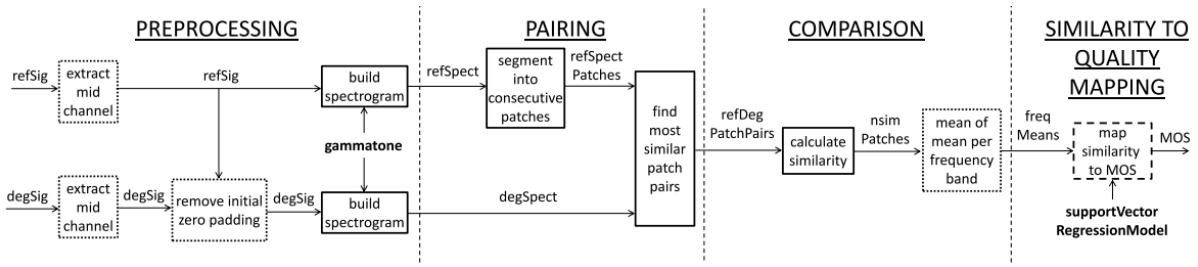

Figure 2.1: A high-level representation of ViSQOLAudio [34].

Chapter 3

Fundamental Concepts

This chapter discusses background information about audio quality assessment methods and its approach in detail. In addition, the chapter gives a detail description about the structure and function of the human auditory processing system.

3.1 Approaches to Audio Quality Assessment

The procedure of evaluating the audio quality is called audio quality assessment. Evaluations are usually expressed as scoring on an ordinal scale for mathematical convenience. There are two categories of assessment: subjective and objective assessment. The audio quality should be evaluated by humans since the term "audio quality" is essentially subjective. Subjective assessment meets the subjective properties of quality and its results are considered truthful. Even though subjective evaluation with human listeners is a ground truth measurement for audio quality, it is time-consuming and expensive to carry out. Due to these reasons, researchers have developed an objective audio assessment method[13].

Audio quality is essentially multivariate. Some quality assessment methods are designed to measure one or more aspects of quality, which can be chosen from the degradation factors mentioned in Chapter one or some intrinsic characteristics of sounds such as naturalness, coarseness, intelligibility, happiness and sadness. Since it is too complicated to identify all quality related factors and measure them, many quality assessment methods focus on an overall quality metric. The way of summarization, which is referred to as "scoring", depends on the definition of the overall quality. An assessment method that judges too many aspects of quality becomes too complex to implement. Usually an assessment with a single metric is more commonly used than a multidimensional metric. Although a single metric assessment method generally fails to give details of the quality,

it is simple to implement and its judgment sufficiently represents listeners' opinions [13].

3.1.1 Subjective Assessment

Subjective assessment is performed as a well-controlled listening test. A group of participants listen to audio clips according to a given metric. Generally, they give scores to represent the quality. It is noted that different persons may have different judgments due to some factors, such as the speaking language used in the test or the contents of the listening materials. Such factors need to be removed or well controlled to decrease the variance in the quality rating. Another way to decrease the variance is to increase the number of participants. When the number of participants is sufficiently large, the total result can be seen as a representation of the true quality.

There are several typical subjective assessment tests. One popular and simple test is the Absolute Category Rating (ACR) test, which is a single metric test and standardized in ITU-T Rec. P.800 [30]. ACR tests are non-intrusive, after each piece of audio file is played, listeners should give their opinions about its quality using a scale shown in Table 3.1. The average of all the ratings is known as Mean Opinion Score (MOS). A disadvantage of ACR tests is that its result may not reflect the true quality of the test signals due to the lack of reference signals. On the presence of a reference signal,

	MOS scale
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 3.1: ACR test grade in MOS scale.

Degradation Category Rating (DCR) is often used, which is also standardized in ITU-T Rec. P.800 [30]. In a DCR test, listeners are provided with a reference signal before they listen to each test signal. They need to rate audio materials using the scale presented

in Table 3.2. The average of all the ratings is referred to as Degradation Mean Opinion Score (DMOS). This category of assessment is denoted as intrusive.

	DMOS scale
Inaudible	5
Audible,but not annoying	4
Slightly annoying	3
Annoying	2
Very annoying	1

Table 3.2: DCR test grade in DMOS scale.

In DCR tests, listeners always take the first signal as reference and evaluate the degraded quality of the second signal. An improved version of DCR is Comparison Category Rating (CCR) test. A CCR test is also a comparison test, in which the listeners similarly evaluate the quality of the second signal relative to the first one using a rating scale described in Table 3.3 and the average is called Comparison Mean Opinion Score (CMOS). However, the order that the reference and the test signals are presented to the listen is random.

	CMOS scale
Much better	3
Better	2
Slightly better	1
About the same	0
Slightly worse	-1
Worse	-2
Much worse	-3

Table 3.3: CCR test grade in CMOS scale.

A subjective test with more refined resolution is "MUlti Stimulus test with Hidden Reference and Anchor" (MUSHRA), which is recommended to assess the intermediate audio quality and is standardized in ITU-R Rec. BS.1534-1 [29]. In this test, a known reference and hidden anchors are included. Listeners listen to the signals in an arbitrary order and for arbitrary times, and rate the quality of test signals from 0 to 100.

3.1.2 Objective Assessment

Objective methods for audio quality assessment have become popular in recent years. Objective methods aim at replacing human subjects with computational models for audio quality estimation. As a result, they provide a quick, cost effective and easily repeatable way of measuring audio quality. Objective methods normally output their results in the form of MOS. Thus, to differentiate between the results obtained by objective and subjective methods ITU-T P.800.1[31] has recommended a mean opinion score terminology. According to this, the MOS obtained by subjective test are denoted by *MOS-LQS(MOS-listening quality subjective)* and MOS obtained by objective tests are denoted by *MOS-LQO(MOS-listening quality objective)*.

The general objective audio quality assessment approaches can be divided into two well-known aspects: *intrusive* and *non-intrusive*. An overview of both methods will be given below and two different test situations can be distinguished in Figure 3.1.

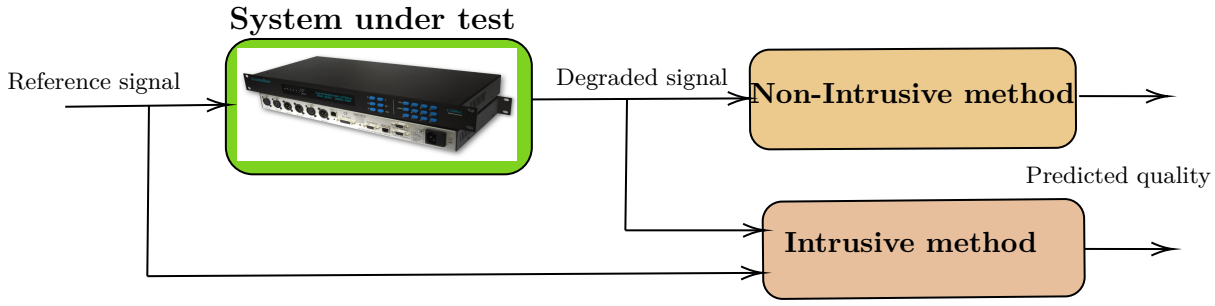

Figure 3.1: Types of objective audio quality assessment.

Intrusive Methods

Intrusive models compare an original test signal with a degraded version that has been processed by a system. Such models have also been termed comparison-based, or full-reference. Both the original and the test signals are available and it is assumed that the original signal itself is of near-perfect quality. As the intrusive method in audio quality assessment is easier with mature theory and practice base that nearly most of researches are based on this model at early times. Most recent intrusive models, which will be discussed in Chapter four, work by transforming both signals using a perceptual model to reproduce some of the key properties of hearing, and then computing distance measures

in the transformed space and using these to estimate MOS-LQO. The framework of most models lies on psychoacoustic and cognitive principles of the human auditory system.

Non-Intrusive Methods

Non-intrusive audio quality estimation is a challenging problem in the sense that a reference, or a clean signal is not available to the computational model. Quality estimates are based solely on the audio signal under test, or on the knowledge of the system under test. Consequently, the results obtained by such models are inferior to those of intrusive models. However, they can be significant for a real-time evaluation of audio and speech quality. These methods are also known as *no-reference* or *single ended* methods.

3.2 The Human Auditory System

The human auditory system includes two parts: ear and central nervous system. The ear is the sensory organ that transforms sound wave from air pressure into nerve impulses. The central nervous system perceives impulses and performs high-level processing. The human ear has three parts: outer ear, middle ear and inner ear. The outer ear gathers and forwards sound waves in the environment. The input sound is filtered as it passes through the middle ear. As the inner ear contains liquid, air pressure, which represents sound, changes into liquid vibrations. The cochlea, which is a part of the inner ear and is surrounded by hair cells, transforms the vibrations into nerve impulses. Those impulses finally reach the central nervous system, which is referred to as the auditory cortex. Figure 3.2 shows the anatomy of the peripheral auditory system. It consists of outer, middle and inner ear which will be explained further below.

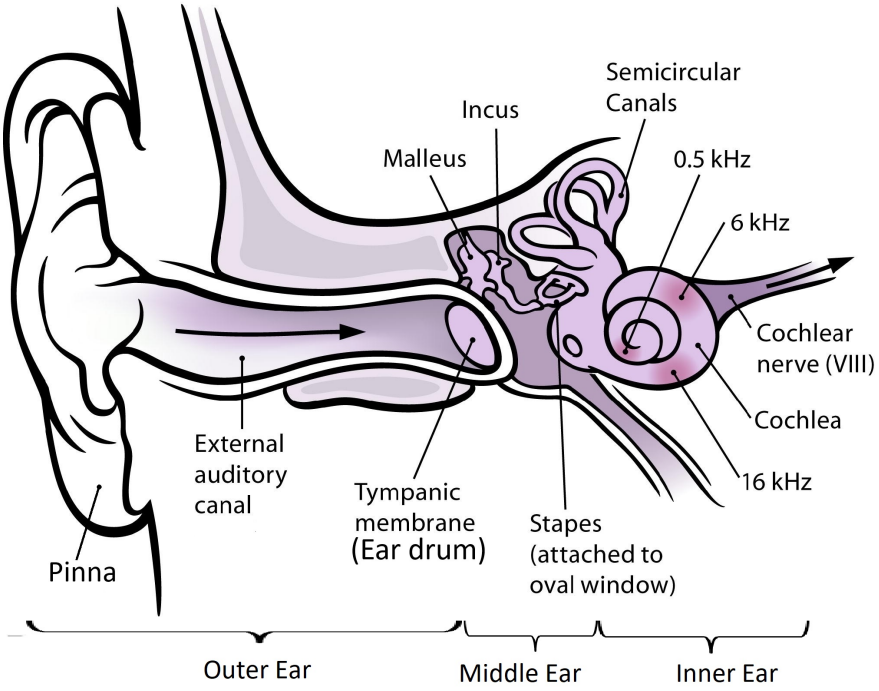

Figure 3.2: Anatomy of the human ear [28].

3.2.1 Outer and Middle Ear

The outer ear is comprised of a pinna and an auditory canal. The pinna reflects sound waves and helps human determine the location of sound source. After being reflected, sound waves propagate along the ear canal and hit the eardrum at the end of the canal. The outer ear filters sound waves and preserves waves in the range from 1 kHz to 9 kHz. The filter frequency response has a peak around 3 kHz. The middle ear is an air-filled hollow cavity, which begins at the eardrum and ends at oval window. This cavity includes three ossicles called the malleus, the incus and the stapes. The main function of the middle ear is to efficiently transfer sounds with a maximal gain of around 1 kHz. The outer and middle ear is modeled as a bandpass filter with a maximal gain at 800 Hz and slops of 20 dB/decade [13].

3.2.2 Inner Ear

Cochlea

The start of the inner ear is an oval window. The inner ear contains a cochlea, which is a coiled tube filled with liquid and has a basilar membrane all along. Air pressure that travels through the middle ear and then is transformed into vibrations of the fluid within the cochlea. As shown in Figure 3.3, there are three fluid-filled spiral tubes in the cochlea. They are the scala vestibuli, the scala tympani and the scala media. The scala vestibuli lies above the scala media, while the scala tympani lies below the scala media. The middle tube "scala media" is also called the cochlea duct and is separated by the basilar membrane from the scala tympani [14].

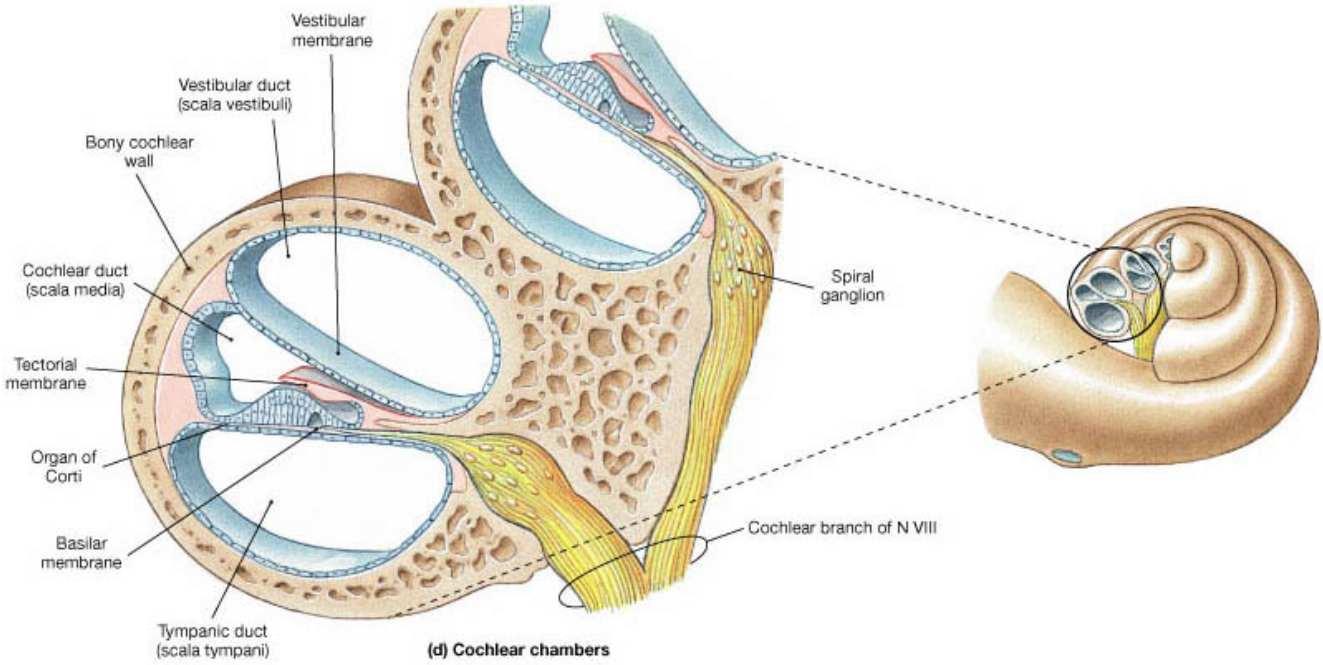

Figure 3.3: Cross section of the cochlear [28].

Basilar Membrane

Basilar membrane has a resonance structure. This means that the membrane has different resonance frequencies at different positions along it. It is shown in experiments that high frequencies resonate near the base of the cochlea (the oval window side) and low frequencies resonate near the apex of the cochlea. This property is resulted from the variance

in the width and stiffness of the membrane. As shown in Figure 3.4, on the oval window side of the cochlea, the membrane is narrower than it is on the other side. The tonotopic mapping of basilar membrane performs essentially spectral analysis. Another discovery is about the relative bandwidth on various positions along the basilar membrane. The ratio between the absolute bandwidth and the center frequency approximately remains a constant along the basilar membrane. This implies that the resolution deceases as the frequency increases [14].

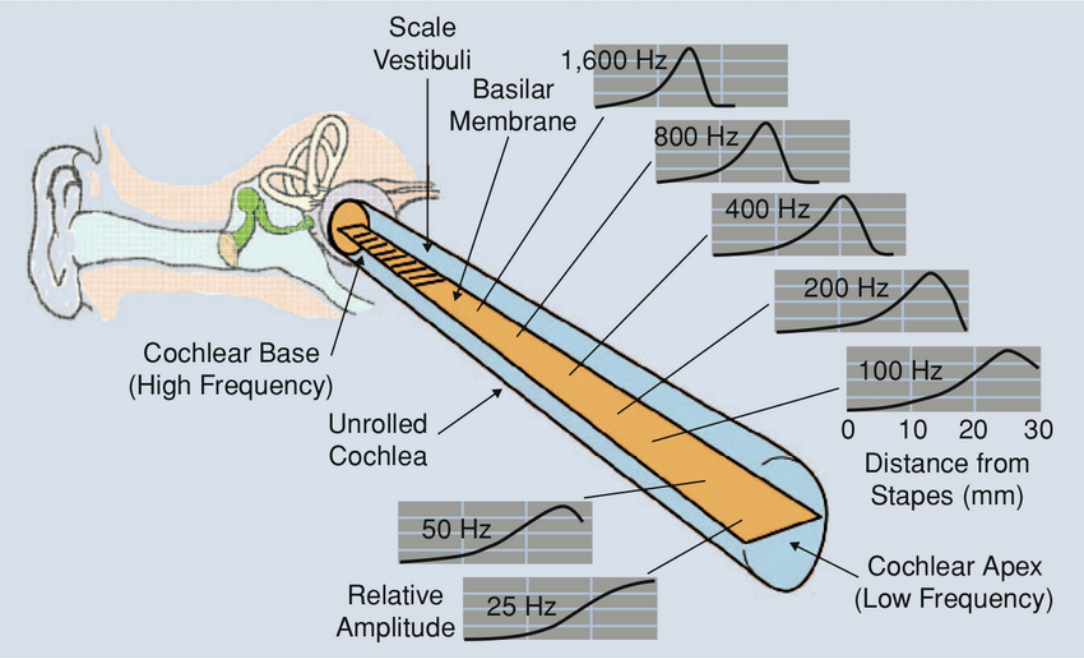

Figure 3.4: Resonance structure of basilar membrane in unrolled cochlear [28].

Hair Cell

As shown in figure 3.5, the organ of corti locates on the basilar membrane. It is a specialized auditory sensory organ that utilizes vibration of the fluids. There are hair cells with stereocilia, which is a mechano sensing organelle and opens an ion channel to let positively charged ion flow into the hair cell as a response to the sound vibration in the fluid. Further, ions that flow into the cell trigger the neural signals. This is how hair cells transform mechanical sound waves into electronic neural signals.

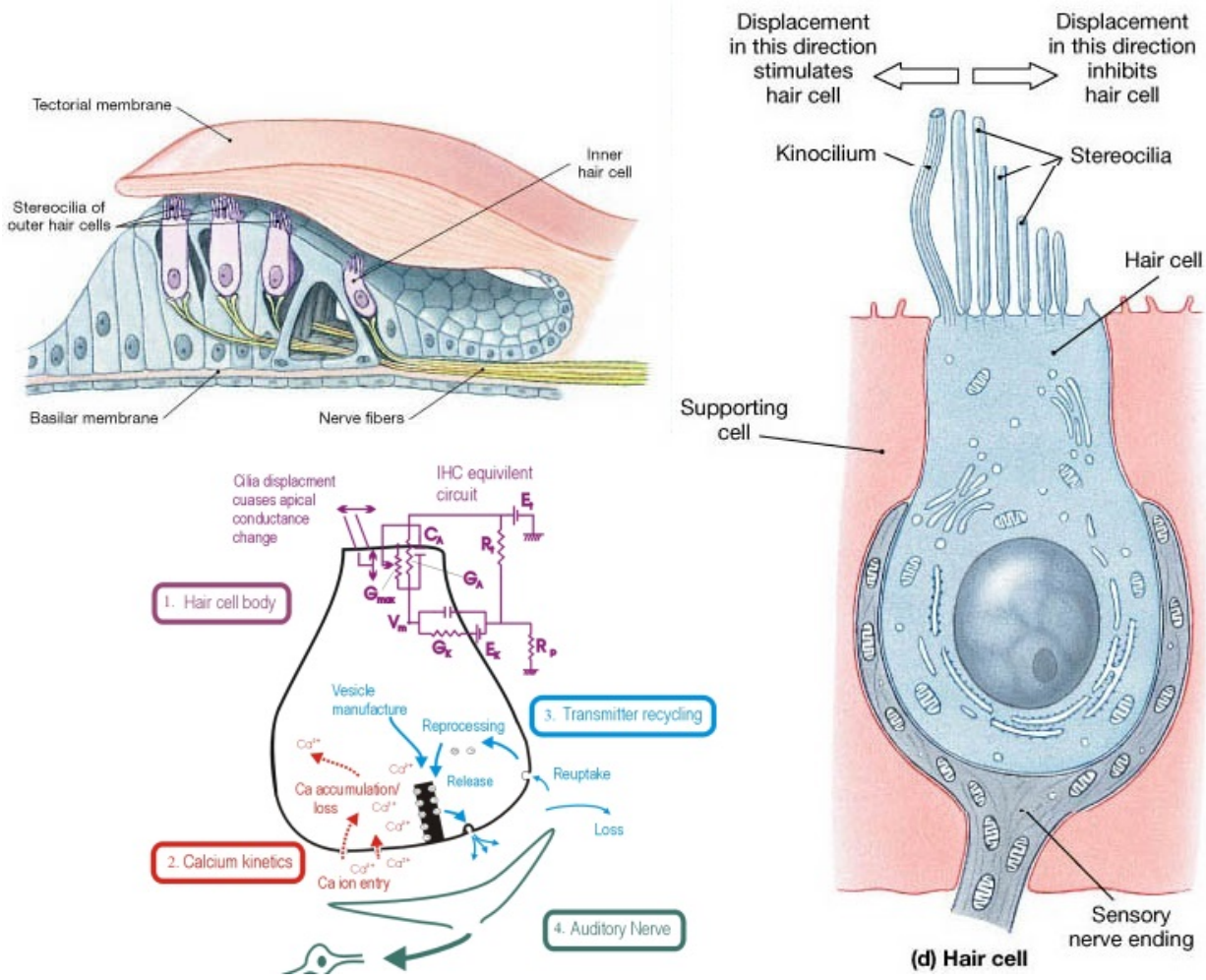

Figure 3.5: Organ of corti, inner hair cell and its equivalent circuit [28].

Chapter 4

Methodology

4.1 Methodology for Designing the New System

In this section, a new method for the objective assessment of perceived audio quality is introduced. It represents an expansion of the latest model, ViSQOLAudio, introduced by*Andrew H. et al.*[34], and is based on a psychoacoustically validated computational model of human auditory signal processing and perception by *Jepsen.M et al.*[19]. To evaluate the audio quality of a given test signal relative to a corresponding high quality reference signal, the auditory model is employed to compute "internal representations" of the signals, which are assimilated in order to account for assumed cognitive aspects [17]. The proposed method, POAQ-CAM, has three main stages *pre-processing*, *auditory processing* and *post-processing*. The approach is we keep the main frame of pre-processing and post-processing of ViSQOLAudio model and acquire the computational auditory model that is applied to auditory processing stage. The overall architecture which is designed for the new method is shown in Figure 4.1. Next, the three main stages of POAQ-CAM will be discussed.

4.1.1 Pre-Processing

Before the reference signal and the test signal are processed by the auditory model, the mid channel, as discussed in the next subsection, is extracted from the reference and degraded signals to consider information from both channels [34]. A possible overall time and level difference of the degraded signal relative to the reference signal has to be eliminated and an initial zero padding at the beginning of the degraded signal should be removed. These deviations are mostly perceptually irrelevant, but could affect the objective quality measure considerably. The steps for the pre-processing stages will be

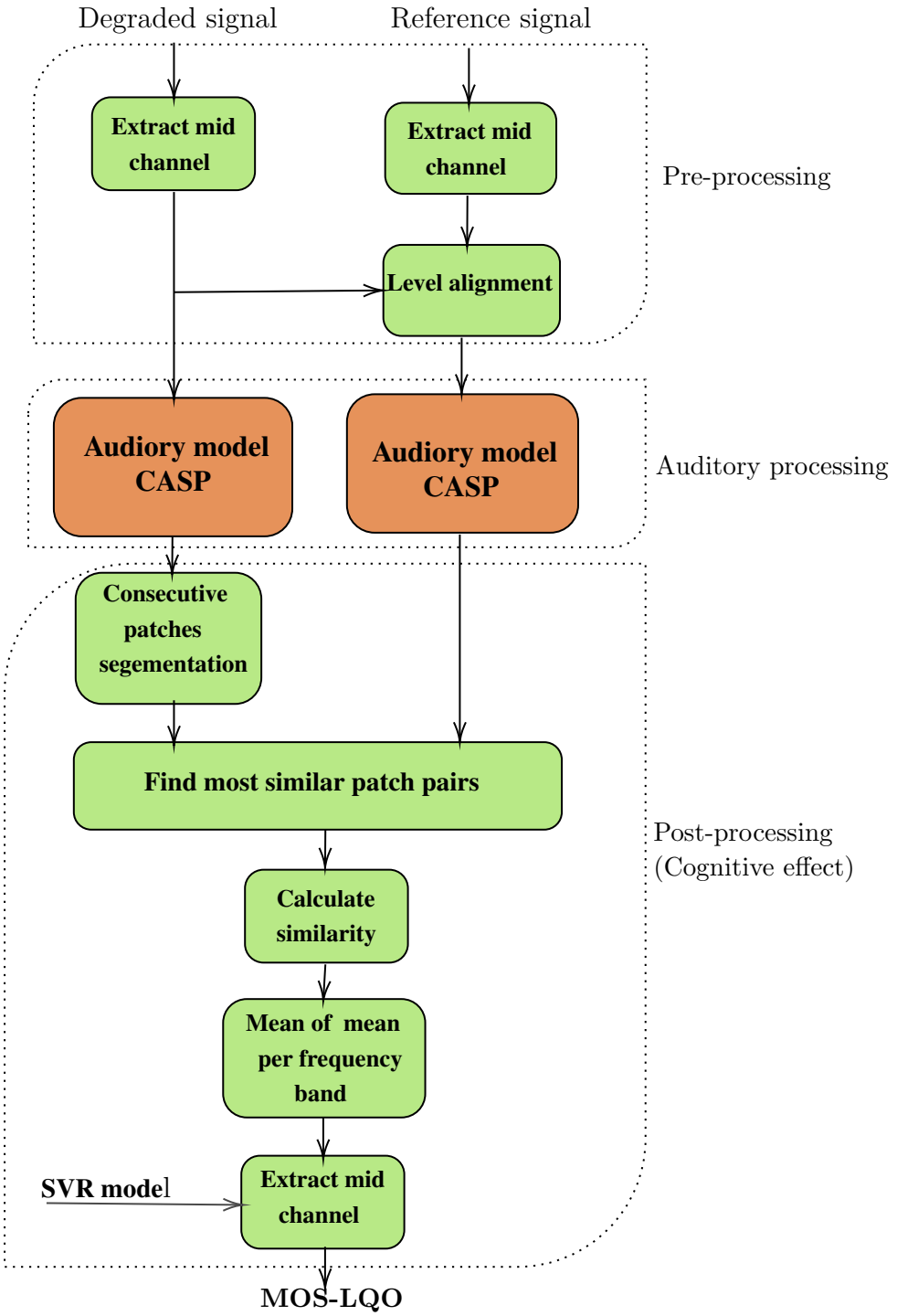

Figure 4.1: System architecture of the proposed POAQ-CAM system.

explained in detail next.

Channel Selection

Subjective studies have shown that perceptual quality of audio output from broadcast

equipment, like codecs, is not uniform for all audios. In audio where the one channel contains more of one instrument than the other channel, an objective model taking information only from one channel may analyze an input signal unrepresentative of the signal heard by the subject. Furthermore, non-expert audio users may upload a stereo audio file containing only one audio signal [34].

Tests revealed that considering the signals from two channels gave more accurate scores than only considering one channel[34]. The two most accurate model stereo configurations came from taking the maximum predicted quality of the left and right channel signals, and the maximum predicted quality of the mid and side channel signals. Further analysis showed that the maximum predicted quality of the mid-side channel pairs almost always came entirely from the mid channel as the side channel contained little information, which meant that it was not necessary to consider the side channel[34]. Besides requiring half of the computational power, using the mid channel signal also alleviated the problems of having different instruments in different audio channels and the problem where users may upload a stereo audio file containing only one audio signal. These results led to the incorporation of the use of the mid channel signal, where

$$mid(y) = (y_{left} + y_{right})/2 \tag{4.1}$$

and y is a left-right stereo input signal.

Removing Initial Zero Padding

When audio is encoded, many popular encoders add a buffer of zero signal samples to the beginning of the degraded signal during the encoding windowing process[26][4]. These additional samples at the beginning of the degraded signal cause misalignment with the uncompressed signal it was encoded from. This misalignment is compensated by using a frequency domain cross- correlation on the Hilbert transformed envelope of the reference and degraded signals to find the correct sample number offset for the degraded signal.

4.1.2 Auditory Processing

To simulate the transformation of acoustic stimuli into neural activity patterns by the human ear, a computational model of human auditory signal processing and perception is applied to the pre-processed pair of reference and test signals. This psychoacoustically motivated model transforms both incoming signals into corresponding 'internal representation'. Figure 4.2 shows a block diagram of the auditory model. The model contains outer and middle ear transformations, a nonlinear basilar membrane processing stages, a hair cell transduction stage, a squaring expansion and an adaptation stage. The model was evaluated in experimental conditions that reflect, to a different degree, effects of compression as well as spectral and temporal resolution in auditory processing [19]. To reduce computational effort and storage consumption for subsequent post-processing steps, POAQ-CAM only considered up to adaptation stages of the CASP model. The stages for the auditory processing will be explained in detail next.

Outer and Middle Ear Transformation

The first stage of the auditory processing is the transformation through the outer and middle ears [19]. A sound pressure waveform produces vibration of the tympanic membrane, which, in turn, induces the vibration of the stapes. Each of these two processes of the peripheral hearing is modeled by means of a linear-phase, 512-point, finite impulse response (FIR) filter. The coefficients of each FIR filter were obtained from empirical frequency responses by applying an inverse Fourier transform[21]. The outer-ear filter was a headphone-to-eardrum transfer function for a specific pair of headphones. The middle-ear filter was derived from human cadaver data and simulates the mechanical impedance change from the outer ear to the middle ear [19]. The combined function has a symmetric bandpass characteristic with a maximum at about 2.5-3KHz and slopes of 20 dB/decade. The output of this stage is assumed to represent the peak velocity of vibration at the stapes as a function of frequency.

The DRNL filterbank

Stapes motion transmits energy to the intra-cochlear fluid, which induces motion of the

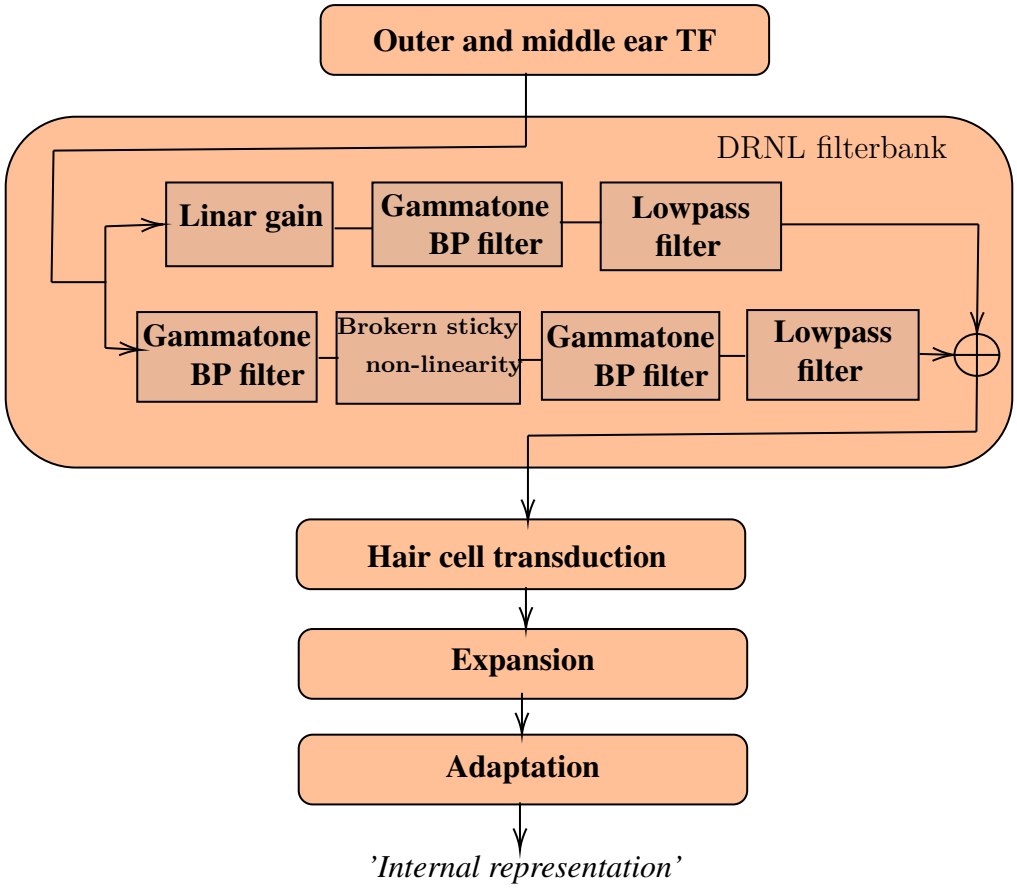

Figure 4.2: Block diagram of the auditory processing adopted from CASP model.

basilar membrane(BM). This process is modeled by a Dual Resonance Non-Linear(DRNL) filter which simulates the velocity of vibration of a given site along the BM in response to a given stapes velocity waveform [21]. This filter includes two parallel bandpass processing paths, a linear one and a compressive nonlinear one, and its output represents the sum of the outputs of the two paths. The structure of the DRNL filter is illustrated in Figure 4.2. The linear path consists of a linear gain function, a gammatone bandpass filter, and a low-pass filter. The nonlinear path consists of a gammatone filter, a compressive function which applies an instantaneous broken-stick nonlinearity, another gammatone filter, and, finally, a lowpass filter. The output of the linear path dominates the sum at high signal levels (above 70 – 80 dB SPL). The nonlinear path behaves linearly at low signal levels (below 30 – 40 dB SPL) and is compressive at medium levels (40 – 70 dB SPL) [2].

The DRNL filter was implemented digitally in the time domain by implementing each of its filters and gains as a digital component. The implementation of each building component was done as follows.

The Gammatone Filters

The gammatone filter has an impulse response of the form [6]

$$h(t) = \begin{cases} kt^{n-1}exp(-2\pi Bt)\cos(2\pi f_c t + \phi), & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{4.2}$$

where n is the order of the filter, B is its bandwidth, f_c is its center frequency, ϕ is its phase and k is a gain.

The DRNL filter uses several cascades of first-order gammatone filters. It is implemented digitally as an infinite impulse response filter as follows:

$$y[i] = a_0 \cdot x[i] + a_1 \cdot x[i-1] - b_1 \cdot y[i-1] - b_2 \cdot y[i-2] \tag{4.3}$$

where $[i]$ refers to the i^{th} sample of the digital signal, x and y are the input and output signals to form the filter respectively. The coefficients are calculated as follows:

$$a_0 = |\frac{1 + b_1\cos\theta - jb_1\sin\theta + b_2\cos(2\theta) - jb_2\sin(2\theta)}{1 + \alpha\cos\theta - j\alpha\sin\theta}| \ ,$$

$a_1 = \alpha \cdot a_0, \quad b_1 = 2\alpha, \quad b_2 = \exp(-2\phi),$

$\theta = 2\pi f_c T_s,$

$\phi = 2\pi B T_s,$

$\alpha = -\exp(-\phi)\cos\theta$

The Low-Pass Filters

The DRNL filter implementation includes several cascades of second-order Butterworth lowpass filters (see figure 4.2). These were implemented as follows:

$$y[i] = C \cdot x[i] + 2 \cdot C \cdot x[i-1] + C \cdot x[i-2] - D \cdot y[i-1] - E \cdot y[n-2]$$

where the coefficients are:

$$C = \frac{1}{1 + \sqrt{2}\cot\theta + \cot^2\theta}$$

$$D = 2 \cdot C(1 - \cot^2\theta)$$

$$E = C(1 - \sqrt{2}\cot\theta + \cot^2\theta)$$

and $\theta = \pi f_c T_s$

where f_c is the $3dB$ down cut-off frequency of the low-pass filter, and dt is the sampling period of the digital signal.

The Linear Gain

The gain in the linear path of the DRNL filter was implemented digitally in the time domain as $y[i] = g \cdot x[i]$, where $[i]$ refers to the i^{th} sample of the digital signal, x and y are the input and output signals from the linear gain stage, respectively.

The Non-Linearity

The time domain digital implementation of the "broken-stick" non-linearity compression function is as follows :

$$y[i] = sign(x[i]) \cdot min(a|x[i]|, b|x[i]|^c)$$

where $[i]$ refers to the i_{th} samples of the digital signal, x and y are the input and output signal from the nonlinearity, and a, b, and c are parameters.

Inner Hair Cell Transduction

The Inner Hair-Cell (IHC) transduction stage in the model roughly simulates the transformation of the mechanical BM oscillations into receptor potentials. This transformation is modeled by half-wave rectification followed by a first order low-pass filter with a cutoff frequency of around $425Hz$. The low-pass filtering preserves the temporal fine structure of the signal at low frequencies and extracts the envelope of the signal at high frequencies [26]. Typically, the low-pass filter is motivated by the loss of phase-locking in the auditory nerve at higher frequencies [4]. Depending on the cut-off frequency of the IHC models, it is possible to control the amount of fine-structure information that is present in higher

frequency channels.

Expansion

The output from IHC transformed into an intensity like representation by applying a squaring expansion. This step is motivated by physiological findings of *Yates et al* [42], which provided evidence for a square-law behavior of rate-versus-level functions of AN fibers near the AN threshold.

Adaptation

The output of the squaring device serves as the input to the adaptation stage of the model which simulates adaptive properties of the auditory periphery. Adaptation refers to dynamic changes in the gain of the system in response to changes in input level. Adaptation has been found physiologically at the level of the Auditory Nerve (AN) [36][35] . In the CASP model [19], the effect of adaptation is realized by a chain of five simple nonlinear circuits, or feedback loops, with different time constants as described by [11][10]. Each circuit consists of a low-pass filter and a division operation. The low-pass filtered output is fed back to the denominator of the divisor element. For a stationary input signal, each loop realizes a square root compression. The output of the series of five loops approaches a logarithmic compression for stationary input signals. For input variations that are rapid compared to the time constants of the lowpass filters, the transformation through the adaptation loops is more linear, leading to an enhancement in fast temporal variations or onsets and offsets at the output of the adaptation loops. The time constants, ranging between 5 and 500 ms, were chosen to account for perceptual forward-masking data[11] In response to signal onsets, the output of the adaptation loops is characterized by a pronounced overshoot. In the study by [10] this overshoot was limited, such that the maximum ratio of the onset response amplitude and steady-state response amplitude was 10 ms.

4.1.3 Post-Processing: Modeling the Cognitive Effects

The stages of simulated auditory processing applied so far represent the auditory perception model of *Jepsone et al.*[19], which was originally designed and optimized to predict detection thresholds from psychoacoustic masking experiments[11][10]. In this thesis, instead of a detected versus not detected decision, it focuses on the similarity comparision between a reference and degraded signal by using a distance metric called the Neurogram Similarity Index Measure (NSIM). A neurogram is analogous to a spectrogram, it presents a pictorial representation of a signal in the time-frequency domains using color to indicate activity intensity. NSIM was developed to evaluate the auditory nerve discharges in a full-reference way by comparing the neurogram for reference speech to the neurogram from degraded speech to predict speech intelligibility [15]. It was inspired and adapted for use in the auditory domain from an image processing technique, structural similarity Index Measure, or SSIM [15], which was created to predict the loss of image quality due to compression artifacts. Next, the four phases of post-processing stage will be discussed.

Building Neurogram

Neurogram, with 38 critical bands, of the reference and degraded signals are created from the adaptation output of auditory processing stages. Its characteristics frequencies (CF) are equally spaced on an equivalent rectangular bandwidth (ERB) scale [12] ranging from $50Hz$ to $16kHz$ that cover both narrowband and wideband audio signals. Figure 4.3 shows, the neurogram image representation of an average power of the adaptation response for a given CF band in the y-axis over time in the x-axis.

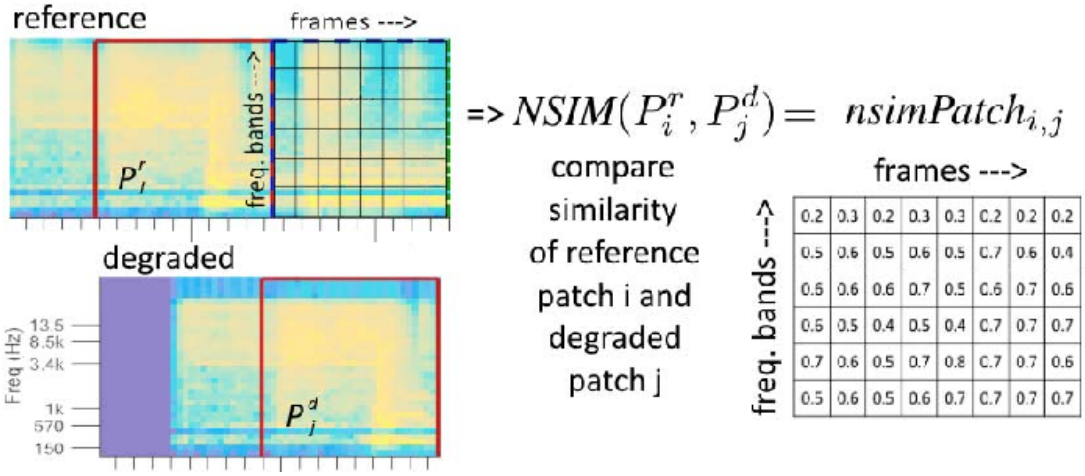

Figure 4.3: The process of creating an NSIM patch by comparing the similarity of a reference and degraded patch pair [34].

Patches Alignment

The reference neurogram is segmented into an ordered set of grids, each 30 frames wide, and with a height equal to the number of critical bands, as shown in Figure 4.3. Each segment is referred to as a patch. The patch alignment process enables compensation for local time misalignment. The goal of the process is to match each reference patch with its most similar corresponding degraded patch, forming a patch pair. A patch pair is denoted (P_i^r, P_j^d), where i is the reference patch index, j is the degraded patch index, P_i^r is a patch from the reference neurogram and P_j^d is a patch from the degraded neurogram [34].

The set of all degraded patches P^d in the degraded neurogram consists of all possible 30 consecutive frames in the degraded neurogram. To find the degraded patch most similar to a reference patch, the process iterates through each possible degraded patch and compares it to the reference patch using the *NSIM*. The degraded patch with the highest similarity measure is selected as the degraded part of the reference degraded patch pair and added to the set of the best patch pairs. The process of finding the most similar degraded patch for a reference patch is described as :

$$s = argmax\overline{NSIM(p_l^r, d)}, d\epsilon p^d \tag{4.4}$$

where p^d is the set of all degraded patches, p_l^r is the reference patch being paired and the over bar is the mean operation. This process is performed for all reference patches, yielding a set of the most similar reference-degraded neurogram patch pairs, *bestPatchPairs* , that will be used during the mapping from patch-pair similarity scores to a MOS-LQO quality score. Before that, we will discuss how the similarity scores used to pair reference and degraded patches are generated [34].

Similarity Comparison

Structural Similarity (SSIM) was originally developed to measure the degradation of compressed JPEG images by comparing the weighted *luminescence*, *contrast* and *structure* of the uncompressed reference image and degraded (compressed) image [41]. Figure 4.3 shows part of a reference and degraded neurogram being compared for similarity. The NSIM of a reference-degraded patch pair, $NSIM(P_i^r, P_j^d)$, is calculated the same way as SSIM index is calculated in [41], but where the luminescence weight $\alpha = 1$, the contrast weight $\beta = 1$, the structure weight $\gamma = 1$, and the regularization constant c_1 and c_3 are 0.01 and 0.03 respectively. Using the windowing method described in [41], a 3x3 Gaussian pixels is the area of interest. The NSIM of a reference and degraded patch is described as:

$$NSIM(P_i^r, P_j^d) = l(P_i^r, P_j^d, c_1) \cdot s(P_i^r, P_j^d, c_3) \tag{4.5}$$

where P_i^r is the reference patch, P_j^d is the degraded patch, l is luminosity and s is structure. Each NSIM value is placed into its respective cell, forming an NSIM patch where a cell represents the similarity between the reference and degraded signals for a given frame and within a given frequency band. As such, patch columns (frames) represent information over time and patch rows (frequency bands) represent information over frequencies.

Mapping Similarity to Quality

The similarity to quality mapping phase inputs the mean frequency band similarity scores of each similarity patch into a support vector regression (SVR) model that outputs a

$MOS - LQO$ value as shown in Figure 4.4 and described as:

$$q = SVR(\frac{1}{M}\sum_{i=1}^{M}\Omega_i) \tag{4.6}$$

Where q is a $MOS - LQO$ value from 1 to 5, M is the number of patches in *bestPatch-Pairs*, Ω is the row (similarity scores across frequency bands) sums of the set of most similar reference – degraded neurogram patch pairs, and SVR is the support vector regression (SVR) mapping function. As shown in Figure 4.4, the row mean, Ω, overall M patches gives a set vectors f, where each f_i is a vector of similarity scores, one for each frequency band. The mean of f is calculated $\bar{f}$ which is input to SVR mapping function that takes a frequency similarity vector as input and outputs a $MOS - LQO$.

The SVR mapping function is an SVR model. The model is a $\upsilon - SVR$ with a radial kernel, where the $\upsilon = 0.6$, $c = 0.4$ and the remaining values are adopted from LIBSVM [7] default. The SVR is trained using $\bar{f}$ as an observation and the degraded audio clip $MOS - LQS$ as the target [34].

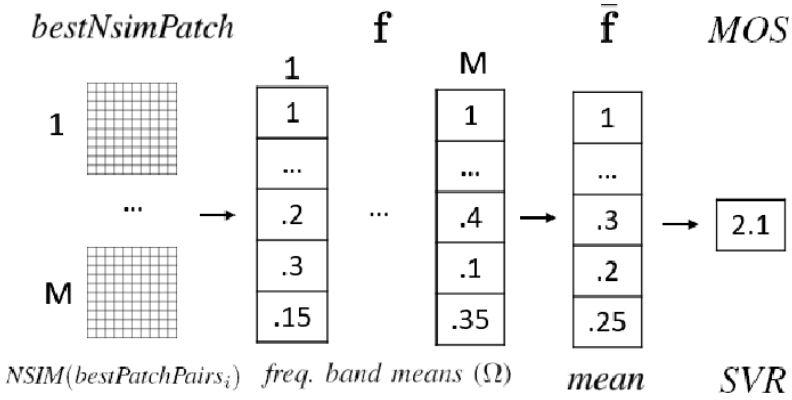

Figure 4.4: The process of generating a MOS-LQO from NSIM patches [34].

4.2 Experiment Methodology

In this section datasets used in the experiments, objective models configuration and evaluation metrics are explained.

4.2.1 Experimental Datasets

Subjective datasets used for metric calibration and testing are a key component in objective model development. Unfortunately, many datasets are not made publicly available; and those that are frequently used do not contain a realistic sample of degradation types targeting a specific application under study, or their limited size does not allow for statistically significant results. MOS scores can vary, based on culture and language, or balance of conditions in a test set, even for tests within the same laboratory. The coverage of the data in terms of variety of conditions and range of perceived quality is usually limited to a range of conditions of interest for a specific research topic. A number of best practice procedures have been set out by the ITU, e.g., the ITU-T P.800 test methodology [9], to ensure statistically reliable results. These cover details such as the number of listeners, environmental conditions, speech sample lengths, and content help to ensure that MOS scores are gathered and interpreted correctly. In this research, for training and testing purpose, a total of 124 subjectively evaluated audio clips and speech samples from three different datasets were used. These datasets include CoreSV [37], TCD-VOiP [23] and ITU-T BS.1387 [18] .

CoreSV

The CoreSV dataset [37] was created to access the quality of the Opus format at 96 kb/s compared to AAC and Ogg Vorbis at 96 kb/s and MP3 at 128 kb/s at a variety of bitrates. There are 40 different samples in total including 5 speech samples and 35 music samples. The music samples contain several solos but mostly excerpts from popular songs across many genres. With six treatments and 40 samples, the dataset contains a total of 240 audio clips[34]. In this research, among the 240 total samples, 60 samples of audio clips were picked randomly as training data, and 10 samples were used for the purpose of testing. The subjective tests used the ABC/Hidden Reference (ABC/HR) methodology, a hidden reference variation of the ABX methodology [22], where subjects played an uncompressed reference and then rated two files: one being the hidden reference and the other being the compressed audio. Ratings were scored on a continuous impairment scale

from 1 (very annoying) to 5 (imperceptible), as described in ITU-T BS.562. This dataset is included in the experiments because it covers a wide range of samples and treatments. Further details on the dataset can be found at [35].

TCD-VOiP

TCD-VoIP is a freely available dataset of degraded speech samples with corresponding subjective opinion scores. The range and level of degradations makes this database a useful resource for the development and testing of speech quality metrics in VoIP and other applications like broadcasting systems. For this database, five types of platform independent degradation were identified and chosen: background noise, intelligible competing speakers, echo effects, amplitude clipping, and choppy speech. From the five types of degradation, 38 speech samples were selected randomly as training data, and 12 samples were used for the purpose of testing. Further details on the dataset can be found at [23].

ITU-T BS.1387

The other dataset used for the development of the audio quality metric in this work is called DB3, which was created for the validation of the ITU BS.1387[32]. This dataset evolved from three listening tests described in (ITU-R,1998a.). The audio reference material consists of 27 stereo sequences processed by six codecs ATRAC(MiniDisc), MPEG layer 2+3, Dolby AC.2 +AC.3 and MPEG AAC alone and in tandem at bit rates from 128 to 256 kbits channel and by adding quantization distortion, harmonic distortion and additive noise. A selection of 84 test signals in total were used in the listening tests. 28 to 33 subjects participated in the listening tests. For this research purposes, 14 training and 3 test data of DB3 were selected.

4.2.2 Objective Models Configuration

POAQ-CAM model used the configuration described in earlier sections. In pre-processing stages, mid channel is extracted, between reference and degraded signal, by removing ini-

tial zero padding, alignment process is performed. Then internal representation of the perceived signal is formed in the auditory processing stage of outer middle ear filtering, DRNL filterbank, Inner hair cell transduction, expansion and adaptation. Finally the coginative effect is analyzed using NSIM and SVR model. The Matlab code for POAQ-CAM can be found at google drive [24].

ViSQOLAudio is a free open source code which is found in [27]. It was tested using the default setting supplied in version 241. Except the auditory processing stages, all stages are incorporated in POAQ-CAM model. PEAQ basic was also considered as a benchmark for the evaluation purpose. PEAQ is freely available as an open source code done by GstPEAQ. GstPEAQ is implemented in plain C as a plugin for the GStreamer framework and it has been successfully used in many internal projects. PEMO-Q was tested with the default settings supplied in version 2.0 of the PEASS Toolkit [31].

4.2.3 Performance Evaluation Metrics

It is recommended that objective models should be assessed at least in terms of their linearity (correlation coefficient) and accuracy (root-mean-square error) [26]. This section defines the metrics used to determine each of these model properties.

Linearity - R

Pearson's correlation coefficient (R) is used to measure the linear relationship between a sequence of objective and subjective quality scores. R is calculated as

$$R = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \tag{4.7}$$

where X_i is the MOS-LQS for audio clip i, Y_i is the MOS-LQO (objective score) for audio clip i, $\bar{X}$ is the mean MOS-LQS, $\bar{Y}$ is the mean objective score, and N is the number of audio clips in the dataset [34].

Accuracy: ϵ- RMSE

The root-mean-square error (RMSE) can be used to describe the absolute prediction error between a sequence of MOS-LQS and objective score values. MOS-LQS values are an average of subjective scores and do not represent variance. The epsilon insensitive root-mean-square (ϵ-RMSE) can be used to describe the prediction error between a sequence of MOS-LQS and objective score values that accounts for variance in the subjective scores . To consider variance, ϵ is first set to the (one-sided) 95 percent confidence interval of the subjective scores that compose a MOS-LQS. An ϵ insensitive prediction error can then be calculated by first predicting an objective score for an audio clip and testing if the score falls within the range of the MOS-LQS ϵ. If it does, the error for that MOS-LQO prediction is set to 0 [34]. ϵ-RMSE is defined as :

$$\epsilon - RMSE = \sqrt{\frac{1}{N-d}\sum_{i=1}^{N} max(0, |X_i - Y_i| - C_i)^2} \tag{4.8}$$

where d is the degree of the polynomial fit and where Ci is the 95 percent confidence interval of the subjective scores for audio clip i. Determining the confidence interval per audio clip is defined:

$$C_i = t(0.05, M_i)\frac{\sigma_i}{\sqrt{M_i}} \tag{4.9}$$

where t is the Student's t-distribution, σ_i is the standard deviation of the subjective scores for the audio clip i, and M_i is the number of subjective scores for the audio clip i [34].

Chapter 5

Results and Discussion

This chapter gives a detailed explanation about Matlab simulation results and performance evaluation of the proposed model, POAQ-CAM.

5.1 Simulation Results and Discussion

Based on the design explained in Chapter four, the obtained simulation results are categorized into two main parts: the pre-processing and the auditory processing, both of which are essential stages of the proposed model, POAQ-CAM.

Pre-processing Stage

One of the main functions of pre-processsing stage is to extract the mid channel from the given stereo audio signal. In Figure 5.1, a two channel (stereo) input (stimuli) and output of mid channel extraction process are shown. Since practically most information is located in the middle of left-right stereo audio signal, from the simulation result, it can be seen that there is no significant difference between the input and output of the mid channel extraction process.

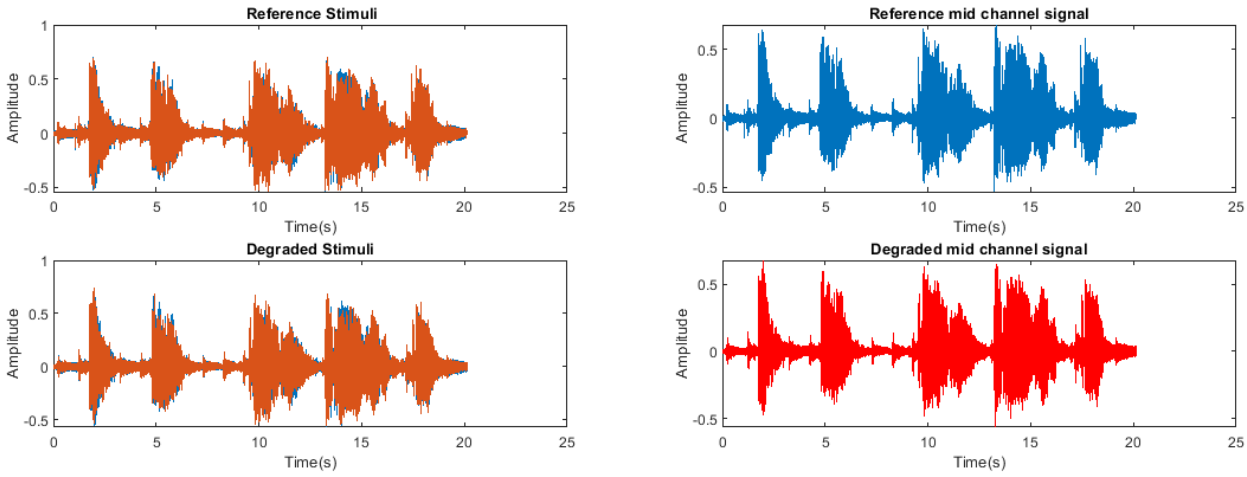

Figure 5.1: Stereo audio input (left panel) and output (right panel) of mid channel extraction process.

Auditory Processing Stage

In auditory processing stage, the input from pre-processed pair of reference and test signals is transformed psychoacoustically into the corresponding internal representation. The first step of auditory processing is the transformation through the Outer and Middle Ears(OME). As explained in Chapter three, the output from OME is the original input plus a resonance. As shown in Figure 5.2, the resonance is modeled as a symmetric bandpass filter with a maximum at about 1 - 2.5 kHz. Also the effect of OME response is clearly verified from the simulation result of time domain and spectrogram outputs shown in Figure 5.3. Interestingly, the input from pre-processing stages are attenuated in the lower and higher register of the spectrum.

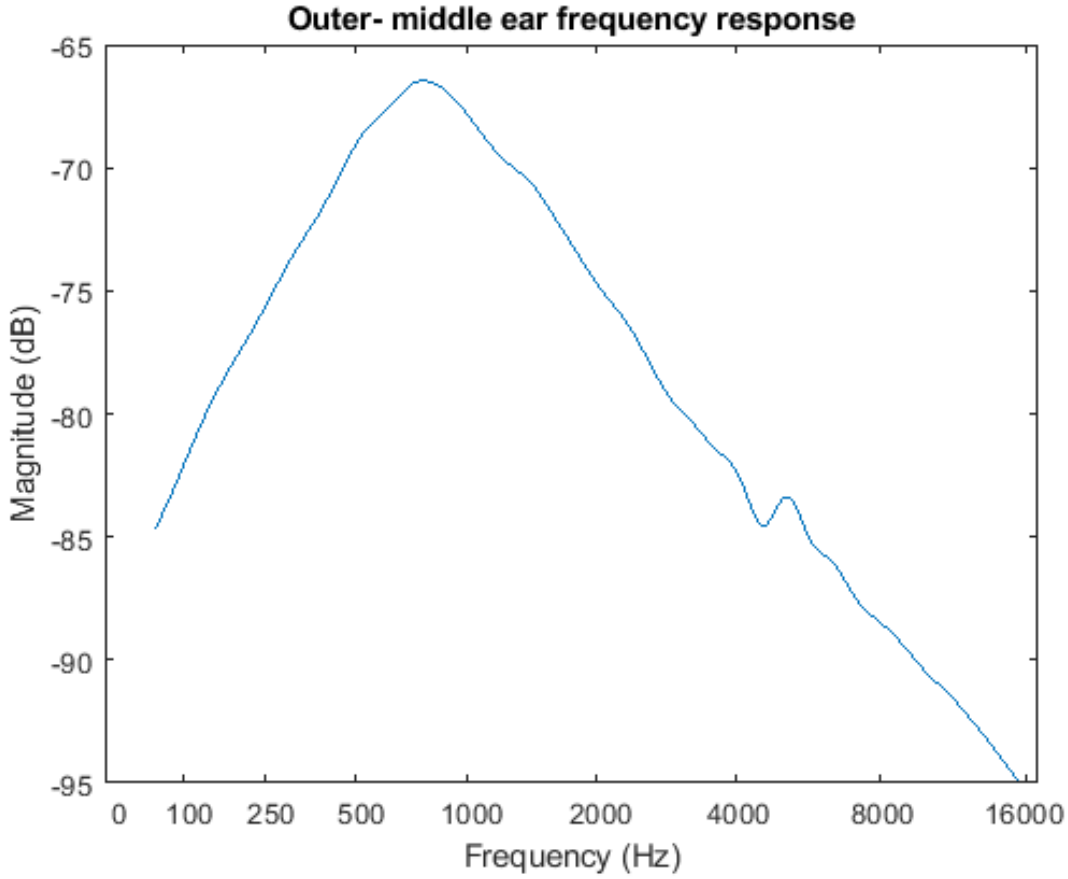

Figure 5.2: Outer and middle ear frequency response.

The other central processing elements of the auditory front end is the separation of incoming acoustic signals into different spectral bands, as it happens in the human inner

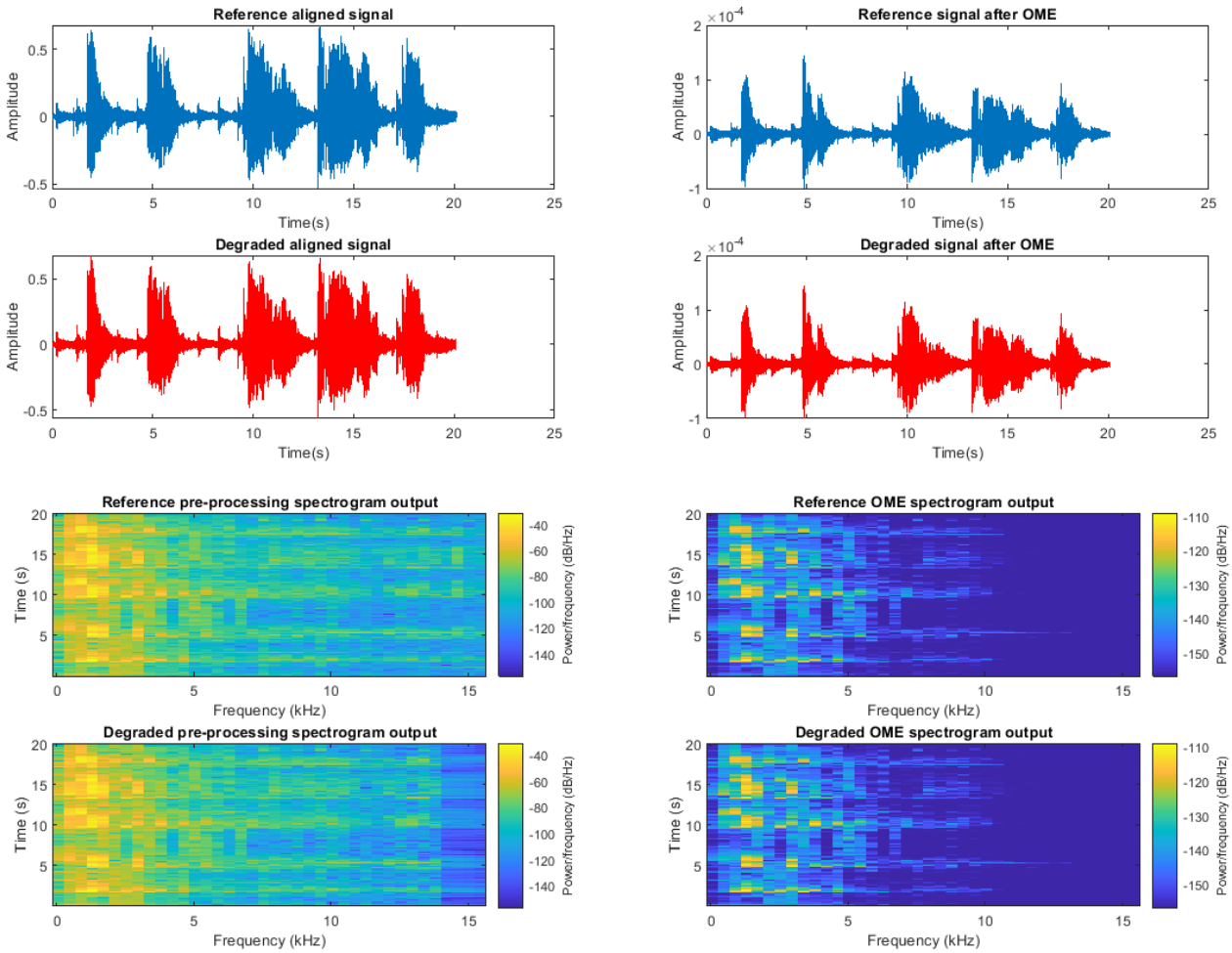

Figure 5.3: Input signal from pre-processing stage (left)and the corresponding output of the outer-middle ear processor(right).

ear. In psychoacoustic modeling, two different approaches have been followed over the years. One is the simulation of a linear filterbank composed of only gammatone filters. The other is, computationally more challenging but at the same time physiologically more plausible, can be realized by a nonlinear DRNL filterbank model. The gammatone filterbank response is illustrated in Figure 5.4. The audio signal shown in (a) is processed through a bank of 16 gammatone filters spaced between 80 Hz and 8kHz. The time domain output of each individual filter is shown in (b). The magnitude response of each gammatone filter is also shown in (c).

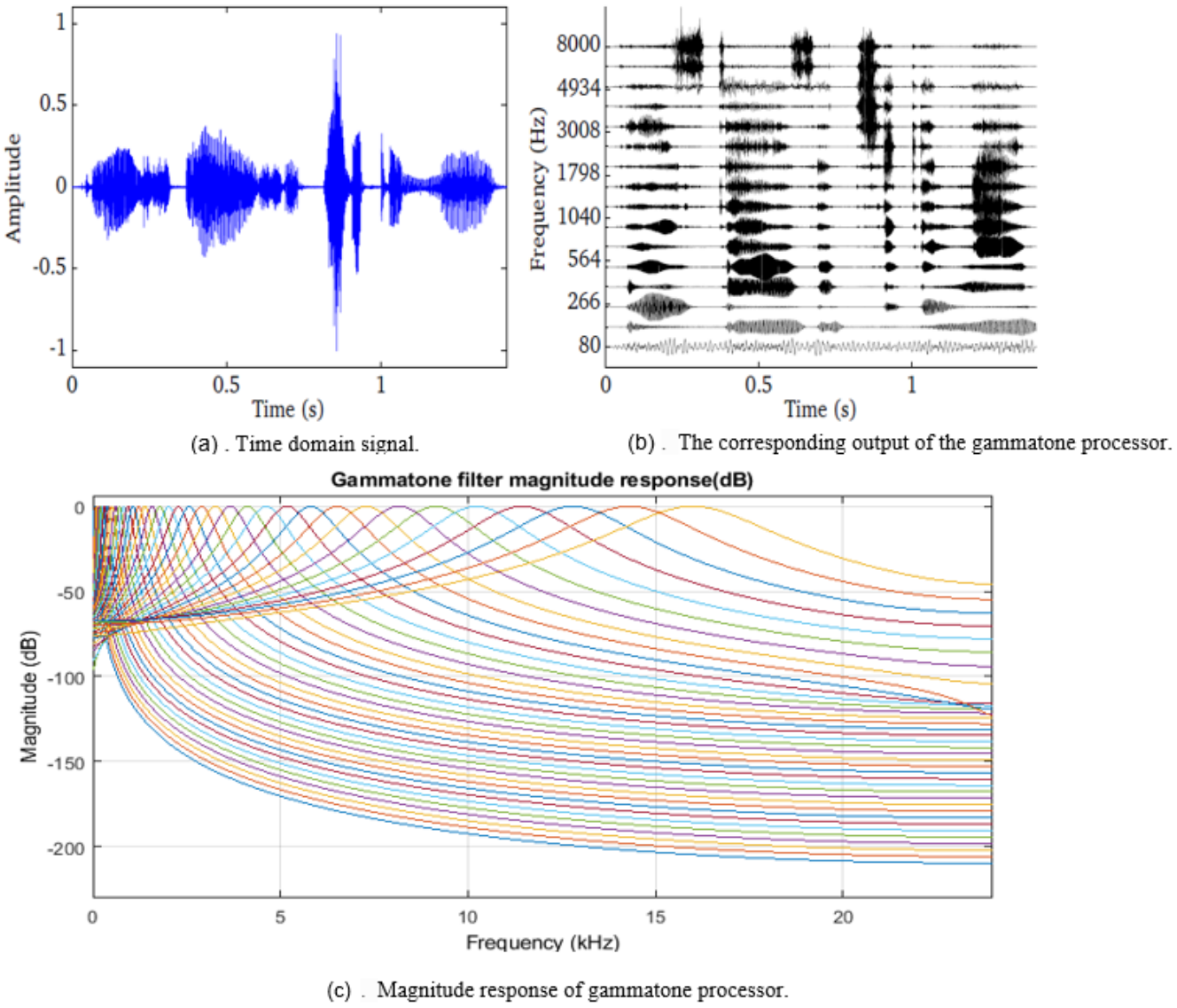

Figure 5.4: Gammatone filterbank processor.

The DRNL filterbank models the nonlinear operation of cochlear, in addition to the frequency selective feature of the basilar membrane handled by the gammatone filters. Figure 5.5 Shows the gammatone processor output (left panel) compared to the output of the DRNL processor (right panel), at 1 kHz center frequency. The input signal is processed using the same pre-processing stage and outer-middle ear filtering before entering both processors for direct comparison. The level difference between the two audio excerpts is reduced in the DRNL response, showing its compressive nature to input level variations.

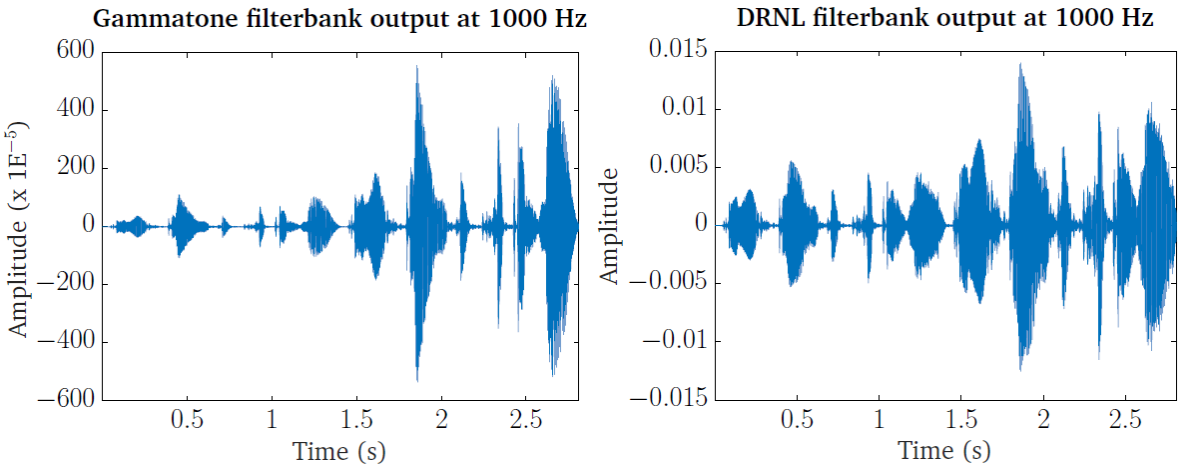

Figure 5.5: Comparison between the outputs of gammatone and DRNL processor.

The effect of the Inner Hair Cell (IHC) process is demonstrated in Figure 5.6. The IHC functionality is simulated by extracting the envelope of the output of individual DRNL filter. Typically, the envelope is extracted by combining half wave rectification and lowpass filtering. The lowpass is motivated by the loss of phase-locking in the auditory nerve at high frequencies. Depending on the cut-off frequency of the model, it is possible to control the amount of fine-structure information that is present in high frequency channels. From the simulation result, the output of this process shows that the individual peaks are resolved in the lowest channel and only the envelope is retained at higher frequencies.

The adaptive response of the auditory nerve fibers is simulated by the adaptation auditory processing unit. The effect of the adaptation process, the exaggeration of rapid variation, is demonstrated in Figure 5.7. This implementation realize the characteristics of the process that IHC input variations which are rapid compared to the time constants are linearly transformed, whereas stationary input signals go through logarithmic compression.

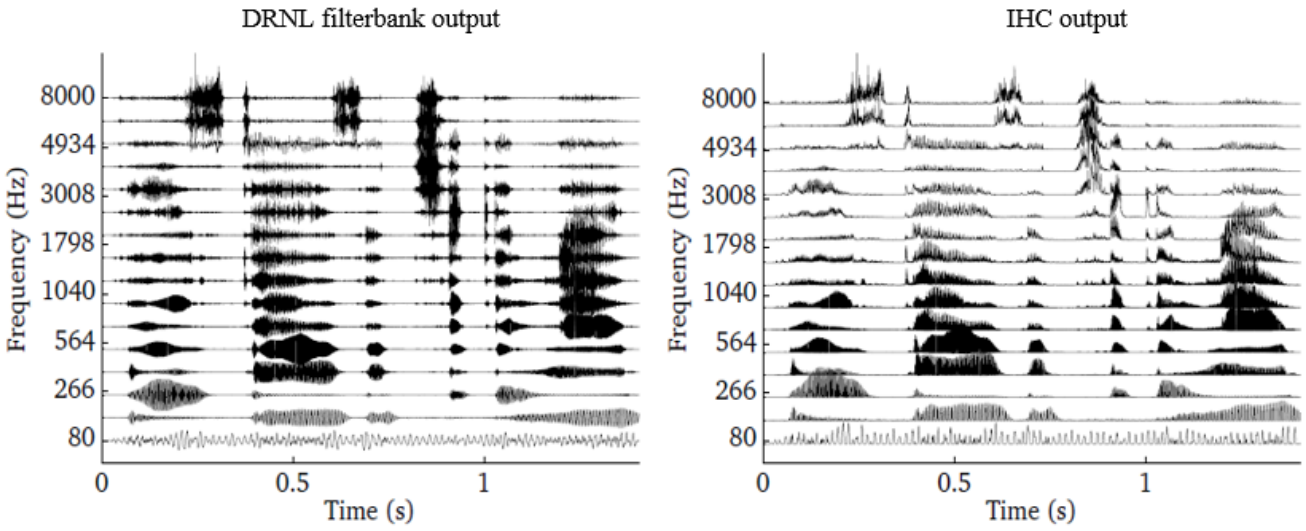

Figure 5.6: Illustration of the envelope extraction processor. BM output (left) and the corresponding IHC model output (right).

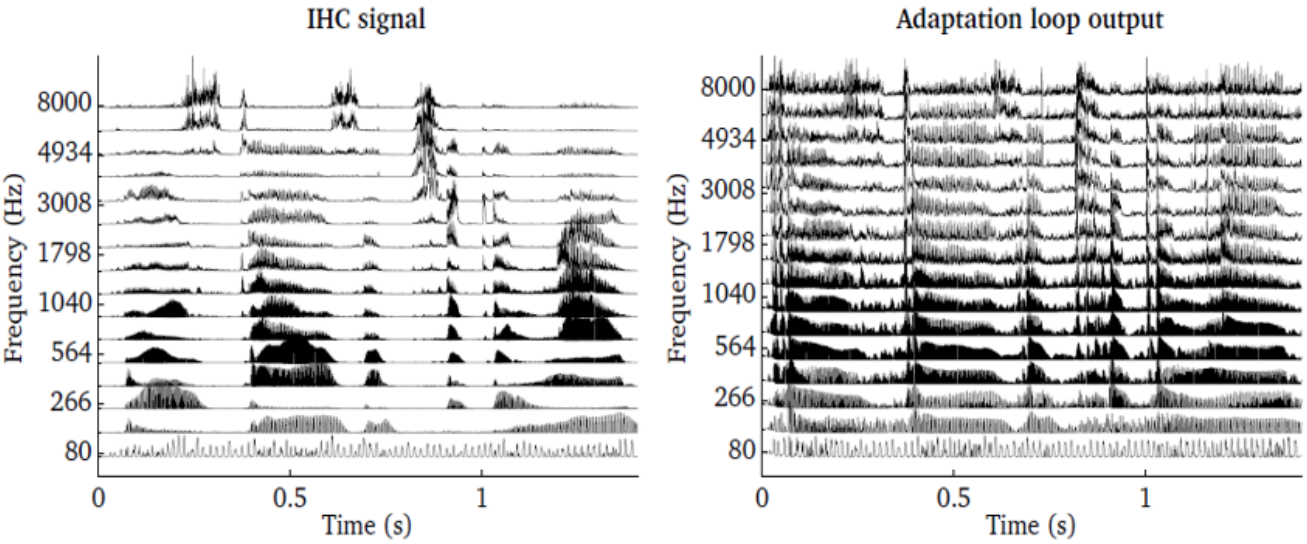

Figure 5.7: Illustration of the adaptation processor. IHC output (left) as the input to the adaptation processor and the corresponding output using(right).

5.2 Performance Evaluation

This section presents the performance evaluation results of applying POAQ-CAM, ViSQOLAudio, PEAQb and PEMO-Q models to each datasets mentioned in Chapter four.

Table 5.1 shows experimental results of objective audio quality assessment models. Each of the models has been trained on datasets to map signal derived attributes to an objective score. Each test has been performed with the same mapping function. A mapping

function is usually trained on several datasets and tested on another datasets. The first column in Table 5.1 shows a collection of audio and speech test samples taken from each of the three datasets for testing purposes. The second column shows subjective difference grade of the respective test samples and the rest column indicates objective difference of POAQ-CAM, ViSQOLAudio, PEAQb and PEMO-Q models.

The results from the tested datasets indicates that PEMO-Q is inaccurate for predicting high audio quality. PEMO-Q also exhibits an unusual pattern of predicting large differences in quality for test samples with the same treatment. This large variation in quality prediction suggests that PEMO-Q is quite sensitive to different kinds of samples content. The prediction for ViSQOLAudio are reasonably accurate for high quality audio clips but inaccurate for low quality clips where ViSQOLAudio seems unable to distinguish high and low audio and speech quality clips. PEAQb is inaccurate when predicting the quality for each of the datasets. PEAQb also has a consistently large variation in its quality predictions for samples with high quality audio.

POAQ-CAM is accurate on all but to the least low quality audio. The variation in POAQ-CAM scores is considerable at low audio quality but reduces to more acceptable levels at greater than 3 MOS-LQS values. This variation suggests that POAQ-CAM becomes less sensitive to different kinds of sample content as perceptual audio quality increases.

Table 5.1 and Figure 5.8 presents the root-mean-square error and correlation coefficient representing the accuracy and linearity of model predictions respectively. Accuracy can be used to describe the absolute prediction error between the objective and subjective score values and linearity used to measure the linear relationship between a sequence of objective and subjective scores. From the evaluation results, comparing POAQ-CAM with PEAQb, PEMO-Q and ViSQOLAudio models, the RMSE is reduced by 43.1%, 48.1% and 22.65% respectively. It shows the proposed audio quality assessment model achieved an improvement in terms of root mean square error over the existing benchmark quality assessment models. When comparing POAQ-CAM with PEAQb and PEMO-Q, the correlation is increased by 6% and 3.9% respectively. But when it compares to

Test samples	SDG	ViSQOLAudio	PEMO-Q	PEAQb	POAQ-CAM
27LSDT.mp3	4.06	4.593670835	1.38966623	1.095413935	3.873929331
Changes.mp4	2.54	4.141467134	1.206274146	1.326164572	3.787106753
clapton.mp4	2.41	4.638547637	1.119599522	1.251013017	3.904484079
ExitMusic.mp4	4.21	3.636798011	1.27222614	1.276542917	3.837152118
girl.mp4	2.58	3.909163089	1.174889177	1.169639315	3.728359659
heytonight.mp4	1.2	3.757443209	1.185287744	1.221876592	3.607407579
Jupiter.mp3	4.1	4.573717479	1.404538827	1.310861563	3.959673878
sandman.mp4	1.36	4.026062656	1.138535031	1.121919194	3.593335678
throughfire.oga	4.68	4.844223064	2.644526636	4.668735147	3.938749005
Waiting.oga	4.03	4.745890209	2.469423772	4.351288007	3.899739575
01 CLIP MK.wav	4.6	4.949224323	5	5.203077103	4.063655981
02CLIP ML.wav	4.5	4.923238595	3.344577322	3.518350473	4.055734912
04 COM FA.wav	3.7	4.846387749	1.137191496	1.252888813	3.696288393
06 COM FG.wav	3.5	4.706377925	1.081835404	1.160567419	3.56631944
06 ECHO MK.wav	3.8	4.838836066	4.139194934	3.918314008	3.923285345
10 ECHO ML.wav	3.7	4.809392557	2.127872516	2.679013646	3.843684761
14 NOISE FA.wav	3	4.598606832	1.109875413	1.093893704	3.231903531
14 NOISE MK.wav	2.8	4.568251298	1.11421792	1.092098619	3.218903946
16 CHOP FG.wav	3	4.874990731	1.609870803	2.444597569	3.945366846
FCODTR1.wav	4.73	4.452531603	4.052240943	4.450728144	3.964045711
GCODCLA.wav	4.08	4.790745665	3.37335013	4.623871097	4.050429184
strauss48 lp35.wav	2.18	2.324653003	1.27912015	2.577533979	3.944069418
guitar48.wav	4.7	4.686782614	1.08950913	1.154902519	3.843339794
sopr48.wav	4.8	4.93287759	1.268941754	1.126491141	4.056527226

Table 5.1: Results of SDG and ODG scores.

Evaluation metrics	ViSQOLAudio	PEMO-Q	PEAQb	POAQ-CAM
$\epsilon - RMSE$	1.2833	1.9060	1.7403	0.9850
$CorrCoef - R$	0.5611	0.5095	0.4957	0.5254

Table 5.2: Result of RMSE and correlation score.

ViSQOLAudio model, it is worth to notice that the correlation is reduced by 5.3%.

Scatter plots show the objective versus subjective scores for each model in Figure 5.9, demonstrating how well each model performed after polynomial regression function is applied. The x-axis of each scatter plot is the MOS-LQS/SDG of an audio clip and the y-axis is the objective quality prediction for the same audio clip. Each solid line is a third order polynomial fit and each dashed line is a first order fit. These scatter plots show that POAQ-CAM and ViSQOLAudio fit well to the subjective scores for each datasets while PEAQb and PEMO-Q consistently underestimates the quality of high

quality audio, where even a third order polynomial regression cannot reasonably account for the prediction.

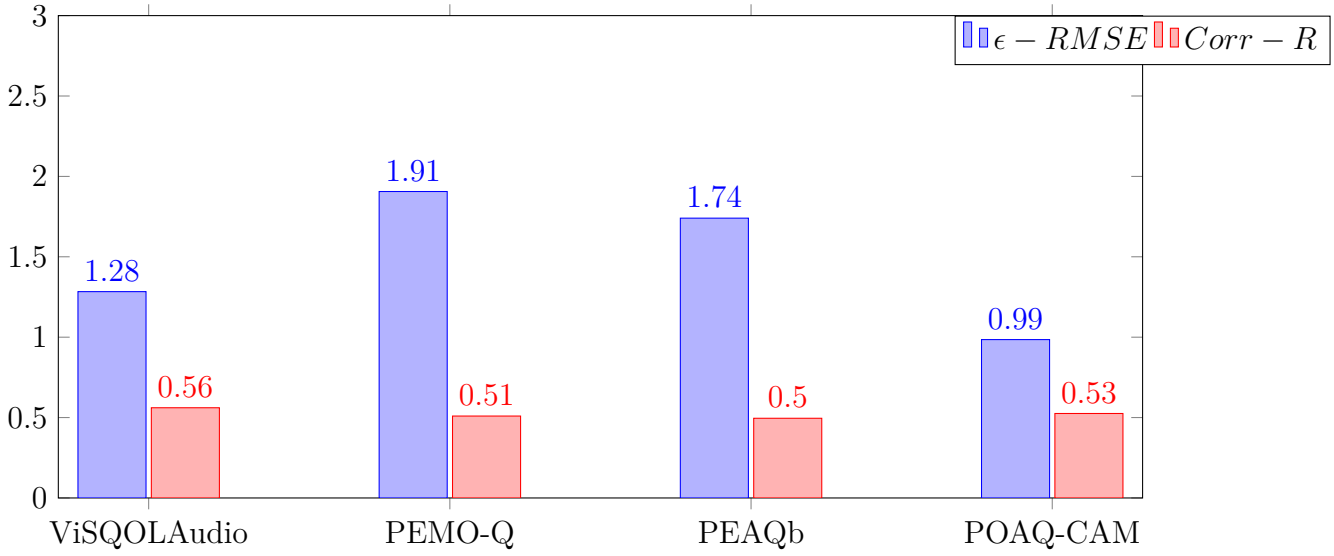

Figure 5.8: Histogram of RMSE and R-correlation score.

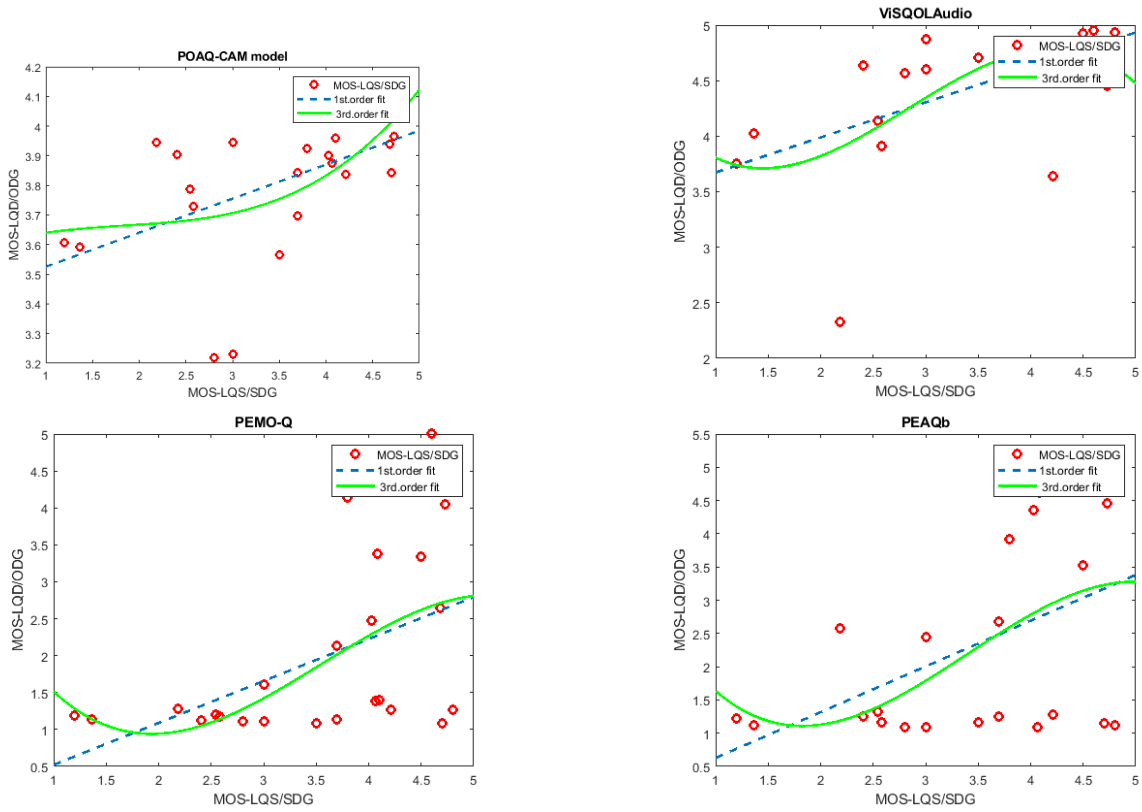

Figure 5.9: Subjective versus objective quality scores. First and third order polynomial regression curves.

Chapter 6

Conclusion and Recommendations

6.1 Conclusion

This thesis work tried to address the problem of audio quality assessment in the digital broadcasting system. A new method has been proposed for prediction of perceived audio quality, POQA-CAM. It is based on a psychoacoustically validated auditory model, CASP, and an expansion of the audio quality assessment model called ViSQOLAudio. It shows that an intrusive objective audio quality assessment method can apply a systematic design based on general purposed auditory models and machine learning tools. This design simulates the procedure that a human rates a signal in a listening test. An auditory model mimics human auditory perception and a machine learning scheme mimics the cogitative procedure that a human perceives audio quality and gives a score accordingly. The proposed method is compared against the latest ViSQOLAudio, PEAQb and PEMO-Q models using a large database of audio and speech subjective listening tests that were originally carried out on behalf International Telecommunication Union (ITU), CoreSV and TCD-VOIP. Compared to the above models, the proposed estimation system has a considerable improvement both in terms of accuracy, measured using root mean square error(RMSE) and linearity measured based on correlation coefficient. The RMSE values of POAQ-CAM is reduced by 22.65%, 43.1% and 48.1% with respect to ViSQOLAudio,PEMO-Q and PEAQb models, while the correlation is increased by 6% and 3.9% compared to PEAQb and PEMO-Q models. It is observed that integrating computational human auditory model into the framework of ViSQOLAudio metric has a positive effect on the accuracy and linearity of the proposed method.

6.2 Recommendations

Subjective datasets used for metric calibration and testing are key components in objective model development. In this thesis, for training and testing purposes subjectively evaluated audio and speech samples were taken from three existing databases. Since subjective tests require a large number of trained human listeners within well controlled listening room and specialized equipment, it is unable to conduct local based subjective test experiments. Because subjective mean opinion scores can vary based on culture and language or balance of conditions in a test set, in the future, it is recommended to establish subjective test setup for specific local research topics. Additionally, in order to elevate the performance and efficiency of POAQ-CAM, it recommended to find a more robust method of machine learning and human auditory model.

Bibliography

[1] Media in Ethiopia . `https://en.wikipedia.org/wiki/Media_in_Ethiopia#Television_and_radio_channels`. [Online; accessed August,2019].

[2] Live broadcast testing. `http://prismsound.com/test_measure/support_subs/apps/live_broadcast_testing.php`, 2008. [Online; accessed August,2019].

[3] John G Beerends and Jan A Stemerdink. A perceptual audio quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, 40(12):963–978, 1992.

[4] Leslie R Bernstein and Constantine Trahiotis. The normalized correlation: Accounting for binaural detection across center frequency. *The Journal of the Acoustical Society of America*, 100(6):3774–3784, 1996.

[5] Karlheinz Brandenburg. Evaluation of quality for audio encoding at low bit rates. In *Audio Engineering Society Convention 82*. Audio Engineering Society, 1987.

[6] Robert P Carlyon and Trevor M Shackleton. Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms? *The Journal of the Acoustical Society of America*, 95(6):3541–3554, 1994.

[7] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. `http://www.csie.ntu.edu.tw//~cjlin/libsvm`, 2011. [Online; accessed August,2019].

[8] Catherine Colomes, Michael Lever, Jean-Bernard Rault, Yves-François Dehery, and Gérard Faucon. A perceptual model applied to audio bit-rate reduction. *Journal of the Audio Engineering Society*, 43(4):233–240, 1995.

[9] Charles D Creusere, Kumar D Kallakuri, and Rahul Vanam. An objective metric of human subjective audio quality optimized for a wide range of audio fidelities. *IEEE transactions on audio, speech, and language processing*, 16(1):129–136, 2007.

[10] Torsten Dau, Birger Kollmeier, and Armin Kohlrausch. Modeling auditory processing of amplitude modulation. ii. spectral and temporal integration. *The Journal of the Acoustical Society of America*, 102(5):2906–2919, 1997.

[11] Torsten Dau, Dirk Püschel, and Armin Kohlrausch. A quantitative model of the "effective"signal processing in the auditory system. i. model structure. *The Journal of the Acoustical Society of America*, 99(6):3615–3622, 1996.

[12] Brian R Glasberg and Brian CJ Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1-2):103–138, 1990.

[13] Ziyuan Guo. Objective audio quality assessment based on spectro-temporal modulation analysis, 2011.

[14] Egger, Katharina. Implementation and evaluation of auditory models for sound localization,2013.

[15] Andrew Hines and Naomi Harte. Speech intelligibility prediction using a neurogram similarity index measure. *Speech Communication*, 54(2):306–320, 2012.

[16] Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte. Visqol: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):13, 2015.

[17] Rainer Huber and Birger 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*, 14(6):1902–1911, 2006.

[18] ITU. ITU-R BS.1387. `http://www.itu.int/ITU-R`, 2001. [Online; accessed August,2019].

[19] Morten L Jepsen, Stephan D Ewert, and Torsten Dau. A computational model of human auditory signal processing and perception. *The Journal of the Acoustical Society of America*, 124(1):422–438, 2008.

[20] Matti Karjalainen. A new auditory model for the evaluation of sound quality of audio systems. In *ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 10, pages 608–611. IEEE, 1985.

[21] Enrique A Lopez-Poveda and Ray Meddis. A human nonlinear cochlear filterbank. *The Journal of the Acoustical Society of America*, 110(6):3107–3118, 2001.

[22] WA Munson and Mark B Gardner. Standardizing auditory tests. *The Journal of the Acoustical Society of America*, 22(5):675–675, 1950.

[23] E.Gillen N. Harte. TCD-VOIP. `http://www.mee.tcd.ie/~sigmedia/Resources/TCD-VoIP`, 2015. [Online; accessed August,2019].

[24] Matlab code for POAQ-CAM. `https://drive.google.com/drive/u/0/folders/1H-Lthv07aEjp6zcpy3qDedKiImx85VuK`. [Online; uploaded August,2019].

[25] Bruno Paillard, Philippe Mabilleau, Sarto Morissette, and Joël Soumagne. Perceval: Perceptual evaluation of the quality of audio signals. *Journal of the Audio Engineering Society*, 40(1/2):21–31, 1992.

[26] AR Palmer and IJ Russell. Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hearing research*, 24(1):1–15, 1986.

[27] Quality of experience for listening and media technologies, qxlab. `https://qxlab.ucd.ie/index.php/2018/05/20/qxlab-wins-research-innovation-award/`. [Online; accessed September,2019].

[28] Tulane University. Auditory transduction. `http://www.tulane.edu/~h0Ward/BrLg/AuditoryTransduction.html`. [Online; accessed August,2019].

[29] ITU-R Recommendation. Itu-r rec. bs.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems. *International Telecommunication Union, Geneva, Switzerland*, 2015.

[30] ITU-T Recommendation. Itu-t rec. p.800: Methods for subjective determination of transmission quality. *International Telecommunication Union, Geneva, Switzerland*, 1996.

[31] ITU-T Recommendation. Itu-t rec. p.800.1: Mean opinion score(mos) terminology. *International Telecommunication Union, Geneva, Switzerland*, 2003.

[32] ITUR Recommendation. 1387: Method for objective measurements of perceived audio quality. *International Telecommunication Union, Geneva, Switzerland*, 2001.

[33] Manfred R Schroeder. Objective measure of certain speech signal degradations based on masking properties of human auditory perception. *Frontiers of speech communication research*, pages 217–229, 1979.

[34] Colm Sloan, Naomi Harte, Damien Kelly, Anil C Kokaram, and Andrew Hines. Objective assessment of perceptual audio quality using visqolaudio. *IEEE Transactions on Broadcasting*, 63(4):693–705, 2017.

[35] RL Smith, ML Brachman, and DA Goodman. Adaptation in the auditory periphery a. *Annals of the New York Academy of Sciences*, 405(1):79–93, 1983.

[36] ROBERT L Smith. Short-term adaptation in single auditory nerve fibers: some poststimulatory effects. *Journal of Neurophysiology*, 40(5):1098–1111, 1977.

[37] CoreSV Team. CoreSV Listening Test. `http://listening.test.coresv.net/results.htm#list2`, 2011. [Online; accessed August,2019].

[38] Thilo Thiede and Ernst Kabot. A new perceptual quality measure for bit-rate reduced audio. In *Audio Engineering Society Convention 100*. Audio Engineering Society, 1996.

[39] Gottfried von Bismarck. Sharpness as an attribute of the timbre of steady sounds. *Acta Acustica united with Acustica*, 30(3):159–172, 1974.

[40] Shihua Wang, Andrew Sekey, and Allen Gersho. An objective measure for predicting subjective quality of speech coders. *IEEE Journal on selected areas in communications*, 10(5):819–829, 1992.

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[42] Graeme K Yates, Ian M Winter, and Donald Robertson. Basilar membrane nonlinearity determines auditory nerve rate-intensity functions and cochlear dynamic range. *Hearing research*, 45(3):203–219, 1990.