# DEVELOPMENT OF AN IMPROVED LOGISTIC MAPPINGFUNCTION FOR OBJECTIVE ASSESSMENT OF QUALITY OF RECEIVED SPEECH OVER MOBILE TELEPHONE NETWORKS

BY

PATRICK OLANIYI OLABISI
B.Eng (Owerri), M.Eng (Akure)
(Matric No. 176403)

A thesis in the Department of Electrical & Electronic Engineering
Submitted to the Faculty of Technology in partial fulfillment of the requirements for the
Award of Degree of

DOCTOR OF PHILOSOPHY
of the
UNIVERSITY OF IBADAN

February 2020.


## ABSTRACT

Users' perspectives approach provides an unbiased assessment of Quality of Service (QoS) of voice offerings in telecommunication networks. This approach is implemented using subjective or objective technique. Though the subjective rating scale is the basis for all ratings, objective technique being computational, can effectively predict quality of degraded speeches. However, existing objective techniques' mapping functions are unable to properly scale speech quality rating. This study was designed to develop a new logistic mapping function that can effectively predict quality of degraded speeches and provide improved quality rating scale.

A speech database consisting of 64 original speeches was developed and transmitted over three mobile telephone networks (A, B and C). Psychoacoustic study was carried out using Zwicker loudness model to evaluate the maximum instantaneous loudness ($N_{max}$) and maximum instantaneous loudness level ($L_{max}$) of original and received speeches. The quality of received speeches relative to the transmitted speech was obtained using $N_{max}$ and $L_{max}$ values. Subjective quality test was carried out on the received speeches based on listening-only technique and Absolute Category Rating (ACR). Objective quality assessment was carried out using Perceptual Evaluation of Speech Quality (PESQ) model, and the scores obtained were mapped using the ITU-T P.862.1 mapping function. The mapped and subjective scores were compared to obtain the correlation coefficient (r), the prediction error ($E_P$), and the Root Mean Square Error (RMSE). A new logistic mapping function for PESQ was developed by optimising the steepness of the logistic S-curve to obtain the growth rate that maximised the range of the quality score. The new mapping function was compared with two international standard mapping functions (ITU-T P.862.1 and Morfitt III–Cotanis). Data were analysed using ANOVA at $\alpha_{0.05}$.

The $N_{max}$ of 46.19 sone and $L_{max}$ of 95.29 phon were obtained for the transmitted speech, and $N_{max}$ of 19.65, 17.13, 16.46 sones and $L_{max}$ of 82.97, 80.98, 80.41 phons were obtained for the received speeches over A, B, and C networks, respectively. The relative quality of the received speeches for the $N_{max}$ and $L_{max}$ were 42.55, 37.08, 35.64% and 87.06, 84.98, 84.37%, respectively. The subjective test of received speeches over networks A, B, and C, resulted in 2.902±0.380, 2.952±0.447 and 2.983±0.612, respectively, while the objective mapped scores were 2.615±0.563, 2.589±0.594 and 2.693±0.730, respectively. Comparing the mapped and subjective scores produced r of 0.854, 0.871, and 0.848, $E_P$ of 0.4264, 0.4724 and 0.4825, and RMSE of 0.4230, 0.4687 and 0.4787, respectively. The optimised steepness resulted in growth rate of 2.2106 and quality coverage of about 1.005 to 4.950. Score coverages of 98.6, 86.8 and 93.7% of the subjective scale were obtained with the developed logistic mapping function, ITU-T P.862.1 and Morfitt III –Cotanis mapping functions, respectively. This indicates a significant improvement over the other two mapping functions.

A new logistic mapping function that enhanced objective technique for users' perspective approach in the assessment of speech quality over mobile telephone networks was developed.



**Keywords**: Psychoacoustic study, Perceptual models, Speech quality assessment, Loudness models.

**Word count**: 477

# DEDICATION

This work is dedicated to a renewed desire, drive and passion to see our great country, Nigeria, begin the walk on the path to technology crafting and break-evens, and the responsible and innovative use of abounding benevolence in human, mineral and capital resources and conducive phenomena of nature that the Almighty God bestowed on this blessed land.

# ACKNOWLEDGEMENT

I hereby acknowledge the fatherly and memorable guidance and supervision given me over the years of my doctoral programme by the astute professor of electrical engineering in the person of Professor Adeboye Olatunbosun. For the contributions of Dr. I. A. Kamil, who served as my Co-Supervisor at the earlier years of my programme, in helping to shape the focus of my research ideas, I use this medium to express my deep heart-felt appreciation to you.

The encouragements of Dr. Ogunjuyigbe and Dr. Zubair and other members of staff in the department are of note in immensely adding fortitude and fiber to helping me forge through to the pulsing point. Indeed, it is a research pulse, knowing there are much more strides to forge and frontiers to attain in this laudable career of knowledge search, discovery and extension.

Dr. Olakanmi has been very wonderful in supporting, guiding, and instructing some of my major steps towards making my works look sensible, formatted and presentable. Thank you in no small measure for your invaluable contributions, directions and supports.

My former head of department at Bells University of Technology, Ota, Dr. Israel Megbowon, directed me to the premier university for my doctoral programme after fruitless attempts and efforts at the programme in three other mega, conventional universities in Nigeria. To him I say God bless you sir, now that the dream is come through.

Finally, to my queen, companion, second pair, and spouse, for her support, encouragement and prayers that strengthened my desire, resolve and struggle at becoming a doctor of philosophy in engineering, I owe you more than mere thanks.

# CERTIFICATION

I certify that this work was carried out by Mr. OLABISI Patrick Olaniyi in the Department of Electrical and Electronic Engineering, University of Ibadan.

_____

Supervisor

Professor A. Olatunbosun

B.Sc. (Hons) (Teesside), M.Sc., Ph.D. (Manchester)

Professor, Department of Electrical and Electronic Engineering

University of Ibadan

# TABLE OF CONTENTS

**LIST OF TABLES**

**LIST OF FIGURES**

**LIST OF ABBREVIATIONS**

AAC – Advanced Audio Coding

ACR – Absolute Category Rating

AIM – Auditory Image Model

AMR – Adaptive Multi Rate

ANIQUE–Auditory Non-Intrusive Quality Estimation

ANOVA – Analysis of Variance

ANSI – American National Standards Institute

AP – Auditory Perceptual

ASD – Auditory Spectrum Distance

BSD – Bark Spectra Distortion,

BM – Basilar Membrane

CB – Critical Bandwidth

CCI – Call Clarity Index

CD – Cepstral Distance

CMOS – Complementary Metal-Oxide Semiconductor

DAM – Diagnostics Accessibility Measure

DFT – Discrete Fourier Transform

DIAL–Diagnostic Instrumental Assessment of Listening quality

DTW – Dynamic Time Warping

E2E – End–to–End

EIH – Ensemble Interval Histogram

ERB – Equivalent Rectangular Bandwidth

ETSI – European Telecommunication Standards Institute

GSM – Global System for Mobile Communication

HPF – High Pass Filter

IHC – Inner Hair Cell

INDSCAL – Individual Difference Scaling

INMD – In-service Non-intrusive Measurement Device

ISDN – Integrated Services Digital Network

ISO–International Standards Organization document

ITU-T –International Telecommunications Union-Telecommunication Sector

KPI –Key Performance Indicators

LPC – Linear Predictive Coefficient

LPF – Low Pass Filter

LSF – Line Spectral Frequencies

MBSD – Modified Bark Spectral Distortion

MDS – Multi-Dimensional Scaling

MFCC – Mel-Frequency Cepstral Coefficients

MNB – Measuring Normalization Blocks

MOS – Mean Opinion Score

MP3 (MPEG – 1 Audio Layer III) – Moving Picture Experts Group

NGN – Next Generation Networks

NMR – Noise to Masking Ratio

NR – Noise Reduction

OBQ – Output-Based Quality

OMC – Operation and Maintenance Center

OPINE–Overall Performance Index model for Network Evaluation

P.AAM – Project – Acoustic Assessment Model

PAMS – Perceptual Analysis Measurement System

PLP – Perceptual Linear Prediction

PSQM – Perceptual Speech Quality Measure

PESQ – Perceptual Evaluation of Speech Quality

PSTN – Public Switched Telephone Network

QoS – Quality of Service

RFC – Random Forest Classifiers

SD – Semantic Difference

SegSNR – Segmented Signal – to – Noise Ratio

SNR – Signal – to – Noise Ratio

SVD – Singular Value Decomposition

SVM – Support Vector Machine

TDH – Time Division Hashing

TOSQA – Telecommunication Objective Speech Quality Assessment

TR – Transmission Rating

UMTS – Universal Mobile Terrestrials System

VAD – Voice Activity Detector

VLSI – Very Large Scale Integration

VoIP – Voice over Internet Protocol

VQ – Vector Quantization

VQoS – Voice Quality of Service

WB-PESQ – Wideband Perceptual Evaluation of Speech Quality

WMA – Windows Media Audio

WSS – Weighted Spectral Slope

# CHAPTER ONE
# INTRODUCTION

## 1.1 Needs to evaluate network service quality

Operations on modern telecommunication networks have become so versatile in view of the dynamic business environment and seemingly insatiable and ever increasing demand by users for new, varied and more effective telecommunication services. Services offered by network operators broadly include multimedia and broadband services and applications of voice, video, mobile television, live streaming of audio or video, music download, data download/upload, data messaging, online games, various internet Protocol applications (Esmailpour and Nasser, 2011).

These services enable us to communicate with people of the world at large and with machines, transact businesses electronically, ensure security of our assets, and correlate our performance and results with those of other intelligent samples and people. We are able to archive important soft documentations in digital repositories, receive or transmit breaking newsand present fresh experiences as they occur in both data and video forms, through telecommunication provisions.Communicating with the outside world, performing remote interactions and conferencing, controllinguse of substances and personnel at various locations, carrying out remote surveillanceand triggering sophisticated war hardwarehave become very possible. Theseinnovative services placed high demand on the provision, utilization and management of network resources (Atif and Zhang, 2014; Ni et al, 2015).

Furthermore, network availability, reliability, pricing and responsiveness to subscribers' complaints are important factors in satisfying users of telecommunication services. It must also be noted that beyond price differentials of telecommunication services, the Quality of Service (QoS) of various servicesoffered by a provider, is a critical differentiating performance factor in the market.

Despite growing range of telecommunication services provided to users, voice service is generally the most patronisedservice provided on telecommunication networks. Voice services include various modes of voice telephony such as voice calls, voice conferencing, voice messaging, voice streaming,and Voice-over-Internet Protocol (VoIP). The quality of voice calls transmitted through telecommunication

networksis of utmost importance to end-users and is paramount in determining the level of effectiveness of QoS offered by telecommunication service providers (Koster et al, 2014).

The continuous monitoring, assessment and analysis of the quality of transmittedspeechover telecommunication networks and optimising it to meet setstandards is of utmost importance. These necessitate studies on improving the quality of voice calls transmitted over telephone networks. Hence the development of approaches and techniques for improving quality of voice calls.

However, before carrying out improvements on the quality of transmitted speeches, the level of degradation suffered by suchspeech signals must first be assessed. Hence, measuring Voice Quality of Service (VQoS) is generally referredto as End–to–End (E2E) speech quality measures implemented with the use of various subjective and objective techniques.

Currently, much emphasis is laid on telecommunication operators measuring and reporting the level of QoS provided to users based on data obtained from network meters at respective nodes and centers. However, very little is mentioned about evaluating QoS from users' perspective. This work assessed the quality of transmitted speeches as perceived by the subscribers through the use of objective (computational) models, whichare in turn correlated with subjective speech quality rating scores.

## 1.2 Overall Telecommunication Quality of Service (QoS) Framework

International standards in ITU-T Rec. G1000 (2001) describe and analyse quality of telecommunication services, and provide the overall telecommunication quality monitoring framework shown in Figure 1.1. The following viewpoints were addressed from boththe customers' and service providers'perspectives:

1. **Customer's QoS Requirement, $QoS_R$**: It is an expression of the measure of E2E QoS required or expected by the customer for a service. It can be represented withvoice quality over the mouth–to–ear model (ITU-T Rec. P.10, 2006)and focused on customer's service needs irrespective of what goes on within the network.
2. **QoS Offered by Network, $QoS_O$**: It consists of QoS conditions offered and clearly specified by the service provider as the basis for Service Level Agreement (SLA), which is also used as planning documents to specify measurement systems within the network.

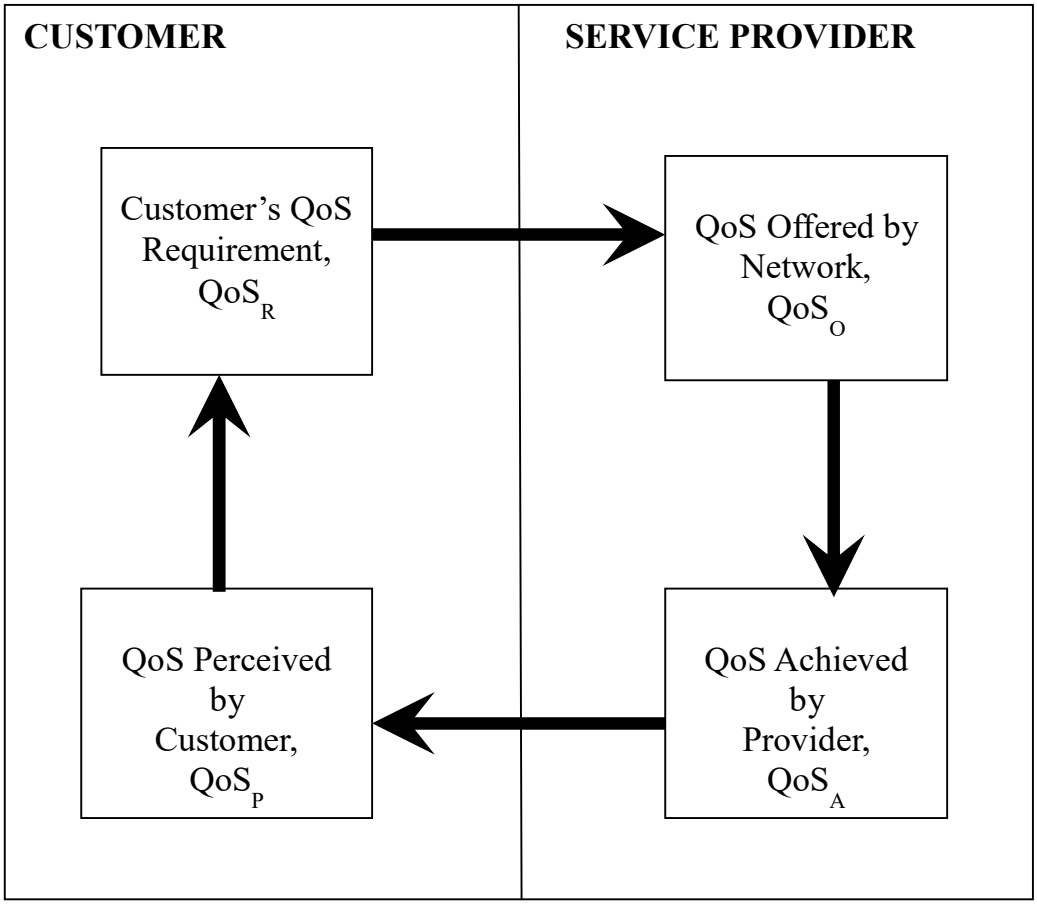



Figure 1.1 Overall Telecommunication QoS Framework (ITU-T Rec. G1000, 2001).

3. **QoS Achieved by Provider, $QoS_A$**: Itstipulates the actual measure of QoS achieved or delivered by the service provider, and used as basis for comparing the level of QoS achieved based on the terms and conditions of service stated in the SLA and for taking corrective or optimisation actions on the network as stated in ITU-T Rec. E.802 (2007).
4. **QoS Perceived by Customer, $QoS_P$**: It is the measure of the QoS received or experienced by users of the service as perceived by them. It is expressed as the degree of satisfaction of the subscribers and quantified with ratings obtained in customer surveys or through psycho-acoustical signaling testing methods. It indicates the level to which the service provider has achieved the required quality statements.

## 1.3 Statement of the Problem

Most efforts at assessing quality of telecommunication services over time have focusedon the quality achieved or delivered by the service provider. Computing this QoS requires use of network performance parameters available at the Operation and Maintenance Center (OMC) of a mobile network, therefore judgment based on it is relative and usually biased. It is opened to manipulation and does not reflect the true views and judgment of users of the services. The satisfaction of network subscribers is therefore not guaranteed despite claims by network operators on provision of quality services.

Our literature review suggests that little efforts were made in most parts of the world at estimating the quality of transmitted speeches over telecommunication networks from users' perspectives.From the survey of previous works, such quality estimation is non-existent in Nigeria in the last 19 years of full-blown wireless mobile cellular telecommunication (GSM and CDMA) in the country. With most available quality estimation reports largely based on network providers' efforts, users are denied opportunity to have first-hand judgment on the quality of speech they send over the telecommunication network to which they subscribed. This is further compounded by poor maintenance culture of these networks. Subscribers are therefore short-changed and unable to make informed decisions against network-based QoS reports supplied by the network operators.

In assessing quality of speeches transmitted over telecommunication networks from users' perspective, objective, automatic, algorithmic, less costly and

computational techniques are adopted for this work, rather than the subjective technique which solely depend on listeners' opinions. The objective approach being mathematically complex, and requiring extensive design and coding of computer-based algorithms have received little research attentions. Where they have been carried out, most objective (instrumental) speech quality techniquesadopted have suffered some algorithmic constraints (Zhang et al, 2013, Shiran and Shallom, 2009,Hu and Loizou, 2008).

The intrusive objective Perceptual Evaluation of Speech Quality (PESQ) model for example, among other constraints, hastheproblem of appropriately rating quality scores of degraded speeches tested on it, based on the bench-marking subjective Mean Opinion Score (MOS) values of the Absolute Category Rating (ACR) scale. This work focused on developing improved functions for more accurately mapping raw objective scores to the standard MOS range. It also studied and developed measures at estimating quality of transmitted speech based on loudness parameters of transmitted speeches.

## 1.4 Aim and Objectives of Research

The aim of this work is to develop improvedmapping functions for quality rating ofobjective perceptual quality assessment of speeches transmitted over mobile wireless cellular networks.

The objectives set forth for the study are as follows:

1. Study key psycho-acoustic parameters of speech which are responsible for determining its quality.
2. Evaluate quality of received speech over wireless cellular networks under distortion conditions for E2E Speech Quality of Service.
3. Study key perceptual speech quality objective assessment models.
4. Develop improvement to thequality score ratings of the objective model for the perceptual evaluation of quality of transmitted speeches.

## 1.5 Research Motivations

1. Evaluation of QoS provided by cellular networks to end-users has been largely network-centric. The perceptualQuality of Service ($QoS_P$) ismissing from QoS monitoring and evaluation activities of network service provisions.

2. The perception of users of services provided by the network is paramount in determining the acceptance of such services.Note that $QoS_P$ closes the quality of service loop.
3. User-perceived QoS verifies adherence to provisions and conditions of SLA, and allows for holding network providers accountable accordingly.
4. Real-time applications, especially voice(conversational and streaming), have stringent delay and delay variation (jitter) constraints, so that the quality of their provision must be assured.

## 1.6 Contribution to Knowledge

1. For Naturalness and Intelligibility of raw speech signals used for the work, speech database was developed locally, though in adherence to guidelines in ITU-T Rec. P.830. This was necessary because all speech databases available for speech quality assessment and processing are all other European and American intonations and none was locally developed before now.
2. With noise reduction and cancellation algorithms built into cellular mobile networks and phones, calculation of psychoacoustic parameters of loudness of speech was proved sufficient for a true picture for speech quality assessment based on comparison of loudness parameters of reference and degraded speeches.
3. With optimised parameters, a logistic mapping function for improved scaling of speech quality scores was developed for the assessment of speech quality over mobile telephone networks based on the international standard model, PESQ.

## 1.7 Organization of the Thesis

In the remaining parts of this thesis, chapter two reviews relevant literatures ontheories, approaches, techniques and models on the topic, as were previously carried out in other researches. Chapter three explains methods and activities that were carried out in this study to implement and validate the research objectives. Chapter four presents results obtained from experimentationscarried out in this study, analyses of these results, graphical and tabular presentations of the results,and discussions of the results. Chapter five concludes the work done so far and makes recommendations for future research efforts.

# CHAPTER TWO

## LITERATURE REVIEW

### 2.1 Overview of Voice Quality Assessment

The International Standards Organization document ISO 9000:2000generally defined quality as: “The totality of features and characteristics of a product or service that bear on its ability to satisfy stated and implied needs.” (OECD, 2006) The International Telecommunications Union (ITU) in ITU-T Rec. E.800(2008) defined the quality of telecommunication services as: “The collective effect of service performance that determines the degree of satisfaction of a user of the service.”

Voice quality implies the clarityand intelligibility of a person’s speech asperceived by a listener. Assessment of the quality of speech in telecommunication applications is carried out in the following areas:

1. evaluation of audio quality in voice systems,
2. selection of speech coders and decoders (codecs), and
3. measurement of voice quality of telecommunication networks.

In voice telecommunication services, Voice Quality Measurement (VQM) greatly espouses the perspectives of end users inthe evaluation of network performance and management. Voice quality portrays a lot of characteristics about the speaker and the network provisions, and also determines intelligibility of the speech. Maintaining voice quality on most telecommunications networks, namely:Public Switched Telephone Networks (PSTN), Cellular Mobile Networks (CMN), andVo-IP, has become very important to both the satisfaction of subscribers and the viability and sustainability of the networks.

The need to ensure acceptable E2EQoS from the calling end-user to the called end-user engendered overall assessment of the characteristic distortions contributedbythe nodes and systems of the telecommunication network(s) on the speech signals that travel through them. Measuring the quality of speeches transmitted over telecommunication networks is both the starting point of improving quality of service and a maintenance tool for quality satisfaction, making the exercise to be very indispensable (Kim and Tarraf, 2004).

Speech quality assessment is very importance in mobile communication networks. Some of the usefulness of speech quality measures noted byJin and Kubichek (1996) include: optimising the design of speech transmission equipment and algorithms, and aiding in the selection of coding algorithms for standardisation.This may also require evaluating the performance of speech coders/decoders (codecs).

Speech quality measurement is majorly subjective and,it signifies the naturalness of how the speech sounds or the effort needed to understand its messageby the receiver (Grancharov et al, 2006). Subjective assessments are generally based on judgments made by subjects (listeners)in standard speech laboratory environment. Standard subjective tests are widely ascertained to be very reliable and accurate methods of assessing users' perception of the quality of speech over a telecommunication network or in speech processing systems.

The alternative measuring approach, the objective voice quality measures, is majorly instrumental and carried out without the use of human listeners. Many techniques of the objective speech quality measures are more recent and still undergoing several and extensive studies. With these, improvements in perceived speech quality have been achieved through applying perceptual methods (Rix et al, 2006).

In this study, various approaches, techniques and measures of both subjective and objective measurement of voice quality which have been developed over the last five decades were reviewed. Shown in Table 2.1 is a general classification of these methods for measuring speech quality (Cote, 2011).

### 2.1.1 Benefits of the Overall QoS Framework

The overall QoS Framework discussed in Section 1.2 provides the following benefits in assessing and resolving QoS problems of telecommunication networks and services:

1. Helps in identifying QoS-related problems in the overall service provision chain;
2. Enables that the problems be quantified from different viewpoints:
   - Customer's viewpoint – use of surveys and subjective tests
   - Service provider's viewpoint – measurements of network performance

3. Ensures that a problem resolved at the provider's end leads to resolving it at the user's end.

Table 2.1.Trends of Speech Quality Assessment Methods (Source: Cote, 2011).

| | **Year** | **1970–'79** | **1980–'89** | **1990 – '99** | **2000 – '09** | **2010–'19** |
|---|---|---|---|---|---|---|
| **Methods** | **Network / Categories** | **PSTN** | **ISDN** | **GSM VoIP** | **UMTS** | **NGN** |
| **Auditory (or Subjective)** | **Analytical** | SD<br>DAM | | | P.835 | |
| | **Utilitarian** | MOS | P.800 | P.830 | P.805 | |
| **Instrumental (or Objective)** | **Parametric** | TR | OPINE<br>II | E-Model | | |
| | **Intrusive** | IS<br>SegSNR | CD | BSD<br>TOSQA<br>PSQM<br>PAMS | P.AAM<br>WB-PESQ<br>PESQ | DIAL<br>POLQA |
| | **Non-intrusive** | | | INMD | CCI P.563<br>ANIQUE | |

## 2.2 Dual Perspectives for Speech Quality Assessment

Networks' perspective and users' perceptual perspective are the two perspectives generally adopted for measuring and assessing the QoS of speech transmitted over mobile wireless networks. These are covered in the overall QoS framework.

### 2.2.1 Speech Quality Measurement from Network Perspective

Generally, the quality of speech transmitted through a telecommunications network could be degraded due to any of the following factors (Rix, 2001; Pocta and Beerends, 2015):

Packet loss; Background noise; Silence; Channel distortions; Frame erasures; Speech processing algorithms like speech encoding (low bit-rate-encoding); Use of assorted algorithms on network systems, like noise suppression algorithms, echo cancellation algorithms, and so on; Delay/Jitter; Echo; and Handset/Terminal Equipment.

Measuring the quality of transmitted speechesovertelecommunication networks bother on issues such as imperfections in voice codecs, noise and distortion on the channel. The approach deployed in measuring the quality of transmitted speech from networks perspective therefore, entails monitoring a number of network performance parameters which include the following (Werner et al, 2003; Kumar and Saini, 2011; Al-Mashouq et al, 2012):

1. Received Signal Strength Indicator (RSSI) also known as Received Signal Level (RxL);
2. Received Signal Quality (RxQual);
3. Carrier-to-Interference (C/I) ratio;
4. Bit Error Rate (BER);
5. Frame Erasure Rate (FER);
6. Handover within a network

RxL is an indicator of signal coverage or spread of signal strength coverage of a mobile wireless network, in –dBm. RxQual provides an estimation of the speech quality carried out by mapping of the bit errors averagedwith respect to a period of

time on the scale of 0 to 7. C/I ratio determineshow much interference suffered by the transmitted speech signal, with highly degraded signal quality having low C/I value.

In using the network parameters to estimate quality of transmitted speech, Al-Mashouq et al (2012) combine four parameters to obtain an estimated quality score from the function given by:

$$q = \sum_{i=1}^{4} a_i w_i \tag{2.1}$$

where, $a_1$ is RxL, $a_2$ is RxQual, $a_3$ is FER, $a_4$ is C/I, and$w_i$ is the weighting factor of the $i^{th}$ parameter.

These network parameters are non-perceptual metrics, and are traditional basisfor measuring and controlling quality of degraded speech (Rohani et al, 2006). Also the location of the user of wireless telecommunication services within the area of coverage of a base station affectsQoS received by the user. Kajackas et al(2004) noted that the user must be within a distance that is optimal and exist under favourable conditions of radio visibility in order to achieve increased chances of successfully communicating and doing so at high QoS. In estimating the impairment of what is known as individual Quality of Service (iQoS), Kajackas and his colleagueslisted three events that should be noted and controlled as the access failure, setup failure, and dropped calls.

Performance of mobile wireless networks is very dynamic due to various atmospheric transmission phenomena, variability of users' needs and unpredictability of users' mobility, system functionalities among other constraints. For the purpose of evaluating and monitoring network performance to determine its level of QoS performance, the characteristics of cellular networks known as the Key Performance Indicators (KPIs) are determined, analysed and reported. For this purpose, specialized data gathering systems and software like protocol analysers, system monitoring protocols and meters/indicators in Operations and Maintenance Centers (OMC) and activities like drive testing are used (Olabisi, 2014).

A set of KPIs listed for service quality are given below (Pareekh, 2010):

I. Service Performance
    1. Round-Trip Time (RTT) Delay (in ms) (800 ms)
    2. Application Throughput (in kbps) (25 kbps)
    3. Call Setup Time (in ms)

II. Network Congestion
    1. Point of Interconnection (POI) Congestion (<0.5%)
III. Connection Establishment (Accessibility)
    1. Call Setup Success Rate (CSSR) (>95%)
    2. Standalone Dedicated Control Channel (SDCCH) Congestion (<1%)
    3. Time Division Hashing (TDH) Congestion (<2%)
IV. Connection Maintenance (Retainability)
    1. Call Drop Rate (CDR) (< 2%)
    2. Worst Affected Cells for Call Drop Rate (<5%)
    3. Connection with Good voice quality (>95%)
V. Service Quality
    1. Prepaid – Prepaid Service Success Rate
    2. Number Portability – Drop Rate
    3. Handover Success Rate
VI. Network Availability
    1. BTSs Accumulated downtime (<2%)
    2. Worst Affected BTSs due to downtime (<2%)

### 2.2.2 Speech Quality Assessment from Perceptual Perspectives

Perceptual methods or approaches at estimating the quality of transmitted speechesover telecommunication networks are either subjective or objective. The original (or reference) speech signal in passing through the network suffers some distortions from the processes of coding and transmission, through network equipment and transmission medium. Figure 2.1 is a schematic diagram of all approaches involved in perceptual speech quality assessment, namely:

1. Subjective Assessment: output (degraded) speeches from telecommunication networks or systems are listened to and rated by listeners as perceived by their auditory systems.
2. Objective Assessment: to predict the quality of transmitted speeches computational models are applied to mimic the perception by human auditory system. There are two types of this approach, namely:
    (i) Intrusive Approach: here both original and degraded speeches are used for computations to predict speech quality;

(ii) Non-intrusive Approach: compute and predict speech quality from extracting psychoacoustic features of degraded (output) speeches and comparing them with those of a reference model.

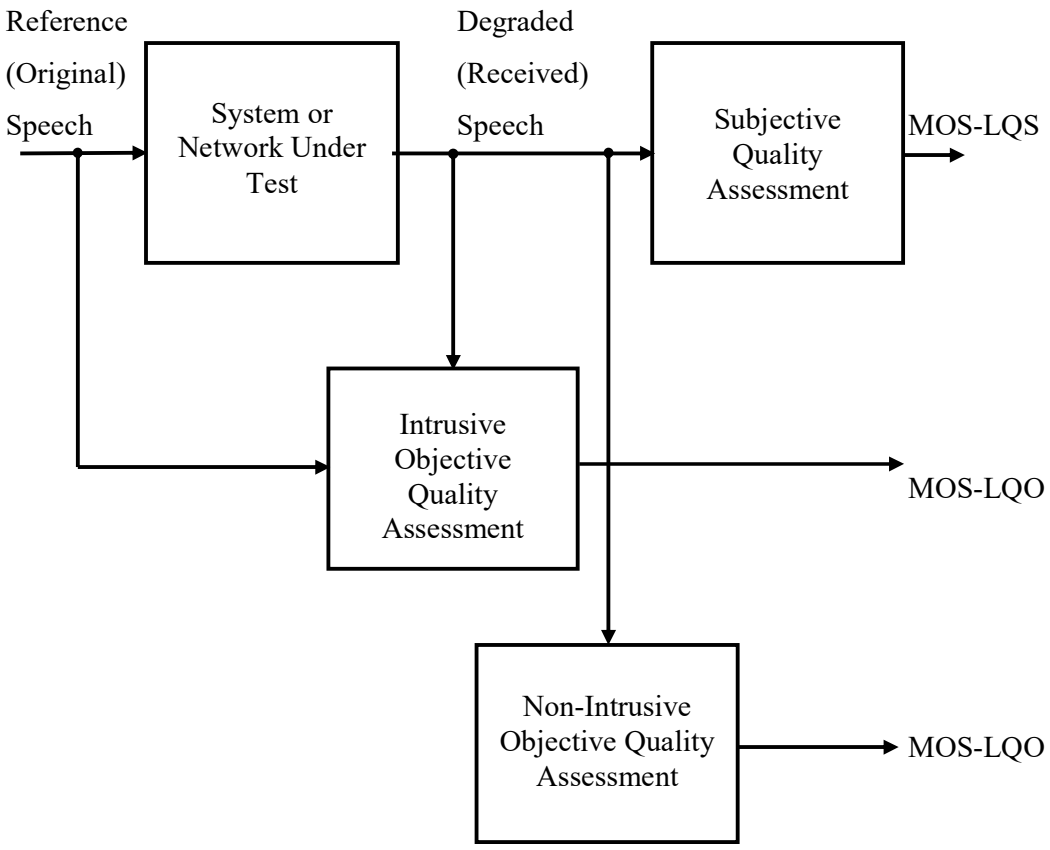



Figure 2.1. Perceptual Speech Quality Assessment Approaches.

*Key*:

MOS-LQS: Mean Opinion Scores – Listening Quality Subjective

MOS-LQO: Mean Opinion Scores – Listening Quality Objective

## 2.3 Subjective Assessment of Speech Quality

Subjective speech quality testing provides overall E2E speech quality score for telecommunication system or network from the perceptual perspective of subscribers, irrespective of the type, design and technology of the network and equipment in use. It is usually carried out in a properly designed speech laboratory whereby, trained subjects (listeners) are asked to listen to hundreds of short live or recorded speech utterances that have been conditioned or processed through different degrading or distorting conditions or obtained as output of a voice system or telecommunication network under test.The speech utterances are played totrained listeners under specialised environmental condition with subjects listening through professional handsets, headsets, or loudspeakers. Listeners rate the speech quality based on their perception and give their opinion about the quality of what they heard with ratings on a five-step scale provided for the test as recommended by ITU-T (Avertisyan and Holub, 2018; Kondo, 2012; Kim and Tarraf, 2004; ITU-T Rec E.800, 2008).

In a subjective test conducted, Bayya and Vis (1996) reported that subjects were required to listen to the following features: background noise, distortion level and overall acceptability. Factors that can affect or shape the opinion of subjects as they listen to rate the quality of speech utterances played back to them include the following (Dimolitsas et al, 1995):

1. The listening level at which speech is played back;
2. The type of filtering applied to the processed or transmitted speech during playback; and
3. The type and quality of listening instrument used for listening – loudspeaker versus telephone handset.

It is widely believed that subjective quality tests give true quality assessment and provide more accurate and reliable means of assessing speech quality (Zhang, 2013 and Rix, 2006). Subjective quality assessment measures help to suggest the need or otherwise for improvements in the network or customer education. Also, large number of listeners is required for subjective test to be statistically valid (ITU-T Rec. E.803, 2011).

Subjective testing of speech quality is in the form of intelligibility test, listening-only test or conversational test. Listening-only test is used in unidirectional speech transmission network test to gather most important quality features.Listeners indicating their opinions of the quality of speeches listened to on a 5-point quality scale known as Absolute Category Rating (ACR) scale shown on Table 2.2. (ITU-T Rec P.830, 1996). The results of listening tests can also be used in assessing two-way connections where degrading effects of sidetone, echo, and delay/jitter are taken into consideration (ITU-T Rec P.800, 1996).

Conversational test is required where important interactive effects manifest in a two-way conversation which listening-only test cannot represent, particularly where conversations of participants are transmitted over different networks and their perception rated(Grancharov et al, 2006). This approach is used for assessing whole link parameters like the network, handsets, sidetones, and impairments, but more expensive and testsfewer degradations than listening tests (Rix, 2004).

All ratings by subjects are collated and averaged for each distortion condition as the Mean Opinion Score (MOS). MOS quality parameter is long-standingand has been adopted for analogue and digital connections and devices, and used for characterizing the quality of telephony equipment and services (Mahdi and Picovici, 2006; Falk and Chan, 2006a). The MOS is also applicable to all types of speech assessment techniques, as shown in Table 2.3.

Subjective test methods developed and used over time include the Threshold Method (TM), whereby subjects directly compare reference speeches with processed speeches and indicate the point of their equality on a regression curve. Semantic Differential (SD) by Osgood et al(1957)was first applied to definethe semantic space in words and was developed to measure what the meaning of objects, events, and concepts or attitudes connote. It used a set of opposing attribute adjectives to measure the reactions people have to stimuli of words and concepts by rating them on bipolar scales defined at contrasting ends. Example byHiese (1970) is:

Good __ __ __ __ __ __ __ Bad
3 2 1 0 1 2 3

Diagnostic Acceptability Measure (DAM) developed by Voiers(1977), is a multi-dimensional scaling method for assessing many different quality features of speech samples. Rated on 20 consistent scales with each scale assigned the

assessment of a specific speech quality feature and read from 0 (negligible) to 100 (extreme) in the following three categories:

1. features that pertain to speech signal (e.g. interruption, rasping);
2. features that pertain to background noise (e.g. hissing, babbling); and

Table 2.2. Absolute Category Rating (ACR) Scale (Source: ITU-T Rec P.830, 1996)

| **MOS** | **Listening quality** | **Equivalent Impairment (Distortion Level)** |
|---|---|---|
| 5 | Excellent | Imperceptible |
| 4 | Good | Perceptible but not annoying |
| 3 | Fair | Slightly annoying |
| 2 | Poor | Annoying |
| 1 | Bad | Very annoying |

Table 2.3. Categories of Mean Opinion Scores (MOS's).

| | Listening Quality (LQ) | Conversational Quality (CQ) | Talking Quality (TQ) |
|---|---|---|---|
| Subjective Quality (S) Testing | MOS-LQSy | MOS-CQSy | MOS-TQSy |
| ObjectiveQuality (O) Testing | MOS-LQOy | MOS-CQOy | MOS-TQOy |
| Estimated Quality (E) | MOS-LQEy | MOS-CQEy | MOS-TQEy |

***Key:***

'y' signifies audio band of quality measurement for the following:

- N is for Narrow Band speech – 300 to 3,400 Hz
- W is for Wide Band speech – 50 to 7,000 Hz, and
- M is for Mixed bandwidth.

3. features that cover both 1 and 2 above (e.g. intelligibility, acceptability).

Noise Reduction (NR) evaluation algorithm addressed background noise in telecommunication systems by increasing the Signal-to-Noise Ratio of transmitted speech. It is evaluated using analytical measurementmethods, which made use of three different 5-point voice quality rating scales for assessing the following quality: the speech signal (speech signal distortion scale), the background noise (background noise intrusiveness scale) and the integral quality (ITU-T Rec. P.835, 2003).

The Multi-Dimensional Scaling (MDS) method is statistical and focused on the perceptual differences in stimuli. The similarities between these stimuli are rated on scales of "very similar" and "not very similar" (Cote, 2011). To rate the quality of speech, a stimulus is placed under test in a previously-defined stimuli space established with the MDS technique using the program INDSCAL (Hall, 2000). A study to evaluateMDS as a technique for finding acoustic characteristics of synthetic speech that would influence how listeners rate the naturalness of aspeech was conducted by Mayo et al(2005).

Despite being adjudged the most accurate and reliable means of assessing speech quality (Kim and Tarraf, 2004), subjective speech quality assessment has the following constraints and short–comings (Dubey and Kumar, 2013; Cote, 2011; Jin and Kubichek, 1996; Grancharov et al, 2006; Mahdi and Picovici, 2006; ITU-T Rec. E.802, 2007):

1. Requires large number of subjects to achieve statistically relevant results – at least 100 interviews per test condition;
2. Very costly to run the test;
3. Very slow, that is, time consuming;
4. Individual human opinion may be overestimated or underestimated, judgements may be misplaced, and performance may be misunderstood;
5. Controlled acoustic environment is required; and
6. Results are highly variable and not easily reproducible.

With these constraints, subjective test measures are unsuitable and fall short of being used in live (real-time) environments and on telecommunication networksor voice processing systems like codecs, and so on.

## 2.4 Objective (Instrumental) Assessment of Speech Quality

Objective methods of assessing speech quality are automatic estimations of perceived speech quality by adopting mathematical models at a level that would equal or be close to ratings obtained from human subjects in subjective assessment, without dependence on human subjects. These methods predict speech quality based on speech signals that are physically measurable.

Objective speech quality models establish relationship between sensational and physical magnitudes. Fechner in 1860 developed the first psychophysical model known as "Weber-Fechner Law" given byCote (2011):

$$S = \alpha \, . \, \ln\left(\frac{\emptyset}{\emptyset_0}\right) \qquad (2.2)$$

where, S is a perceived intensity of sensation, αis the constant of proportionality, $\emptyset$= physical parameter and $\emptyset_0$= perception threshold.

Over time, developing accurate objective measures for speech quality assessment,researchers have based their works on constructing models that extensively utilize characteristics of the human perceptual auditory system along withtheirperceptual/hardware equivalence. Though they are computationally very intensive, objective quality measures are used for monitoring speech quality on telecommunication networks based on the experience of end users. These have the ultimatepurpose of optimising networks and speech processing systems for better quality performance, increased capacity and cost effectiveness (Rix et al, 2006).

General characteristics of objective speech quality assessment techniques are as follows (Grancharov et al, 2006; Falk and Chan, 2006a):

1. They offer automatic voice quality assessment on communication systems;
2. They are based on complex mathematical models;
3. They utilize algorithms for computing MOS value from a small piece of the particular speech;
4. They can be coded and computerized;

5. They are extensively used to support results of subjective tests;
6. They are less costly means of implementing signals quality assessment;
7. They are used on real-time basis to continuously measure quality of speech on live telecommunication networks or voice processing systems.

Objective speech quality models are majorly classified into parameter-based, signal-based and packet-layer models. Parameter-based (or parametric) models mostly make use of estimates of network properties to predict quality of speech communication and are usually carried out during network planning before the implementation (Rix, 2004). Such models include loudness rating model, opinion model, and E model. They make little or no use of perceptual techniques, but predict speech quality scores using measured properties of network ranging from type of codec, to echo, bit rate, delay, speech levels (loudness), packet loss, noise, and other measured network characteristics requiring full network or system characterization (Koster et al, 2014; Rix et al, 2006).

In parameter-based measurements, the following cellular network transmission parameters were identified (Werner et al, 2004):

1. RxQual – involves averaging the BER and mapping it to the log RxQual.
2. AMR Mode – specified the mode of the Adaptive Multi-Rate (AMR) codec used in GSM networks. Lower modes imply higher error correction capabilities but lower speech quality.
3. FER – the Frame Erasure Rate.
4. MnMxLFER – the Mean of Maximum Lengths of Erased Frames.

Signal-based models make use ofthe degraded signals from an existing telecommunication networks or speech processing systems in evaluating the quality of speech through them. A number of different signal-based methods are based on models of speech production or likelihood, while others explore areas of perception like noise loudness (Rix et al, 2006). Objective signal-based models are either intrusive or non-intrusive and are briefly discussed below.

### 2.4.1 Intrusive Objective Assessment of Speech Quality

The Intrusive models, also referred to as double-ended or input-output-based models, extract key auditory features through a process of perceptual transform from

an original or “clean” speech known as the reference speech. It also extracts those of the processed or degraded version obtained from a transmission network or speech processing system. The features of these speeches are compared, and the amount of their deviation is used to compute an estimated MOS, from which the level of quality of the degraded speech with reference to the original speech is determined.

Intrusive speech quality techniques make use of perceptual models for assessing speech quality. The earlier ones include the Masked-error model, which was proposed by Schroeder in 1979 and extended by Brandenburg in 1987 (Cote, 2011). In estimating how audible coding noiseare in codecs, simple masking techniques were used to obtain the mean of the Noise to Masking Ratio (NMR), such that the difference on the time frame between reference and distorted speeches was counted to be noise (Rix, 2004).

The waveform-comparison algorithm models namely the SNR and the Segmented SNR(SSNR) techniques require low computational algorithms and so are simple to implement but do not correlate well with subjective assessment results at the face of comparison of diverse distortions (Kondo, 2012; Grancharov et al, 2006).

The SNR-based techniques do not sufficiently provide a prediction of speech quality in modern telecommunication networks. It led to more complex assessment measures being developed. Some of these assessment measures were discussed in (Liu et al, 2006; Bayya and Vis,1996), and they include the Cepstral Distance (CD) which compares two smoothed spectra in the cepstral domain, Log Spectral Distance (SD) which obtains the log difference of the power spectra of the original and the degraded speeches, the Weighted Spectral Slope (WSS) which is based on weighted differences between spectral slopes of 36 overlapping frequency bands, and the Auditory Spectrum Distance (ASD) which compares representations of the original and the degraded speeches in terms of the audible time (in ms), pitch (in Bark) and amplitude (in dB).

Perceptual-domain models require that psychoacoustic processes are utilized to transform both the original and degraded speech signals in accordance with the auditory system. This transformation follows the psychoacoustic model for calculating loudness which was developed by Zwicker and Fastl (Cote, 2011). Combination of perceptual transforms and simulation of the cognitive processes in human auditory cortex are used to obtain an estimated integral quality of speech.

Perceptual-domain models include Bark Spectral Distortion (BSD) whose perceptual transformation emulates auditory phenomena of critical integration in the cochlea and the loudness compression.The calculated distortion is the square of the Euclidean distance that exists between the respective speeches. Modified Bark Spectral Distortion (MBSD) incorporated a noise-masking threshold so as to make a difference between audible and inaudible distortions. Perceptual Speech Quality Measure (PSQM) was developed in 1994 and published asITU-T Rec. P.861(1996). It compares a coded signal to a source signal by mimicking sound perception and judgement processes of humans.

Other perceptual-domain models include Measuring Normalization Blocks (MNB), which was developed by Voran, S (1999, Parts I & II) and based primarily on useful parameters of objective estimators of perceived speech quality namely: perceptual transformation and distance measures reflecting the magnitude of the perceived distance between two perceptually transformed signals.

PSQM as one of the four most important intrusive objective modelsdescribes audible network or system errors as distances between the original and the degraded speeches and obtained the total errors from the error spread. Other models including Perceptual Analysis Measurement System (PAMS) developed by British Telecoms (BT) in 1998 (Mohamed, 2003). It was primarily designed to correct network properties of linear filtering and bulk delay variations which made previous models unsuitable for E2E speech quality assessment in telecommunication networks (Rix and Hollier, 2000). The state-of-the-art Perceptual Evaluation Speech Quality (PESQ), and the most recent Perceptual Objective Listening Quality Analysis (POLQA) (Voznak, et al, 2013) are discussed below. For these methods, the quality of coded or transmitted speeches is determined based on differences in the internal representation to calculate the noise disturbances in time and frequency.

Perceptual Evaluation of Speech Quality (PESQ) was deployed only for transmitted speeches on telephone networks and standardized as ITU-T Rec. P.862(2001) It is a more robust model for speech quality measurement than the PSQM. It incorporates the perceptual transformation feature of PSQM99 model as well as the time-alignment algorithm routine of PAMS model. The wideband version of PESQ (WB-PESQ) was developed in 2005, and currently the most widely used perceptual model (Cote, 2011). The most recent of these models, the Perceptual Objective Listening Quality Analysis (POLQA), is aimed at correcting alignment

defects in PESQ model and at predicting integral speech transmission quality for all types of telecommunication networks in Next-Generation Networks (NGN).

### 2.4.2 Non-Intrusive Objective Assessment of Speech Quality

Non-intrusive objective assessment speech quality models, also referred to as output-based, no-reference or single-ended quality models, do not require the original speech as reference signal for any comparison of features with those of the degraded speech signal. It isappropriatefor assessing speech quality at network end points or at any point or node on a telecommunication network (Dubey and Kumar, 2013).

The constraints of intrusive models that necessitate developing non-intrusive models include:

1. Alignment of the original and degraded speech signals is very difficult due to the variable delay it introduces, and this result in decrease in the accuracy of intrusive models.
2. In some applications where it may be impossible for the reference signal to be present, for example in network monitoring, intrusive models may not be so appropriate.

Non-intrusive models are implemented based on two approaches: the a-priori based and the source based approaches.

#### 2.4.2.1 The a-priori approach

This approach has two major model types,namely: those based on the use of codebook of speech features and those based on the speech production system.In the first type of models, a codebook which contains characterised parameters of a set of known distortions is developed. This set of parameters is used to train learning machines, to deducehow these parameters relate with the perceived speech quality. Machine learning techniques used in speech quality assessment include the Gaussian Mixture Models (GMM), Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forest Classifiers (RFC) (Falk and Chan, 2006b).

For quality models related to speech production vocal system, there is the possibility that the degraded signal was produced by the human vocal system,and this is a focal point for consideration. Speech features derived from speech production mechanism are actually extracted from the speech signal, and are integrated into a

quality scale. Example of this type of models is the Vocal Tract (VT) model developed by (Gray et al, 2000), and published inITU-T Rec. P.563 (2004). This model operates based on three principles, namely:

1. Derivationof several parameters from the degraded speech signal concerning voice production mechanism;
2. Reconstruction of a reference signal from the degraded speech signal; and
3. Detection of specific distortions in the degraded speech by comparingit with the pseudo reference signal to predict the speech quality of the communication system or network.

#### 2.4.2.2 The source-based approach

This approach also depends on learning machine techniques which maps an artificial reference signal with the parametric features obtained from the degraded speech signal,which are Perceptual Linear Prediction (PLP) coefficients, Mel Frequency Cepstral Coefficients (MFCC) features, and Line Spectral Frequencies (LSF) features. Euclidean or cepstral distance is therefore calculated between the artificial and the degraded signals to estimate the level of degradation suffered by the degraded signal and the speech quality is thereby estimated.

## 2.5 Voice Production System

Physiological process of human speech production is made of three major functional units: generation of air pressure happens in the lung, regulation of vibration happens in the larynx, and control of resonators carried out at the nasal and oral cavities, which are briefly considered below (Zhang, 2016; Honda, 2008). The lung in a process known as *respiration* produces the sound energy source, the larynx in a process known as *phonation* converts the energy from the air pressure into audible sound (voice production), and the articulators in a process known as *articulation*converts the sound into intelligent speech.

The human speech production system shown in Figure 2.2displays various sections and components responsible for voice source, air flow manipulation known as phonation and the formation or articulation of proper phoneme of speech.The block form in Figure 2.3 shows the respiration, phonation and articulation stages in speech production, while Figure 2.4 shows the signals waveform representation at these stages including the final speech output.

Outward push of air from the lung resulting from operations of the respiratory system provides the source of the human speech. The air stream passing through the vocal cord in the larynx is controlled by a set of laryngeal muscles, and in the process vibrates the vocal folds. The larynx is fixed on the trachea. In it are found the major vocal components known as the vocal folds shown in Figure 2.5. At this point, the air stream from the lung is converted into a form of quasi-periodic buzzing pulse sound made of intermittent airflow through the opening and closing operations of the vocal folds vibrated by the air stream as shown in Figure 2.6.

Voice production is controlled by the brain with the aid of various nerve connections and signals. Such signals include the signal for the movement of the muscles of the voice box (lungs), that is, the motor nerves, which comes from the Recurrent Laryngeal Nerve (RLN) and the Superior Laryngeal Nerve (SLN), and the signal from the voice box mechanism for feeling, that is, sensory nerves which flow through sensory paths of RLN and SLN (voiceproblem.org).

The rate at which vocal folds vibrate is known as the fundamental frequency ($F_0$) of the human voice. It is the pitch of voice expressed in Hertz (Hz). This is relative to the size of the vocal folds, which is responsible for wide difference in the average fundamental frequencies of the sexes and age groups: male adults – 100 Hz, female adults – 200 Hz, children – 300 Hz (Cote, 2011).

Transmitted through the vocal tract are the sound fundamental frequency ($F_0$) and its harmonics, with energy peaks of the sound frequency spectrum concentratingin formants, particularly the first three formants. The first formant, $F_1$, is the formant of the lowest frequency and has most of the energy concentrated in it. The range of the human voice frequency is majorly 100 – 7,000 Hz, but because the human voice can generate complex intelligible sounds like whistle, hisses, hum, clicks, and so on, may increase to about 50 – 14,000 Hz (Cote, 2011).

Considering the acoustics of the vocal tract, linear wave motion in the vocal tract is formulated after the law of continuity and Newton's law, and given by O'Shaughnessy (2000):

$$\frac{1}{\rho c^2}\frac{\partial p}{\partial t} + \operatorname{div} \mathbf{v} = 0 \tag{2.3}$$

$$\rho \frac{\partial \mathbf{v}}{\partial t} + \operatorname{grad} p = 0 \tag{2.4}$$

where $t$ is time, $p$ is sound pressure, $c$is the speed ofsound in air ($c$ = 340 m/s), **v** is vector velocity of air particle in the vocal tract and $\rho$ is the density of air in the tube ($\rho$ = 1.2 mg/cm$^3$).

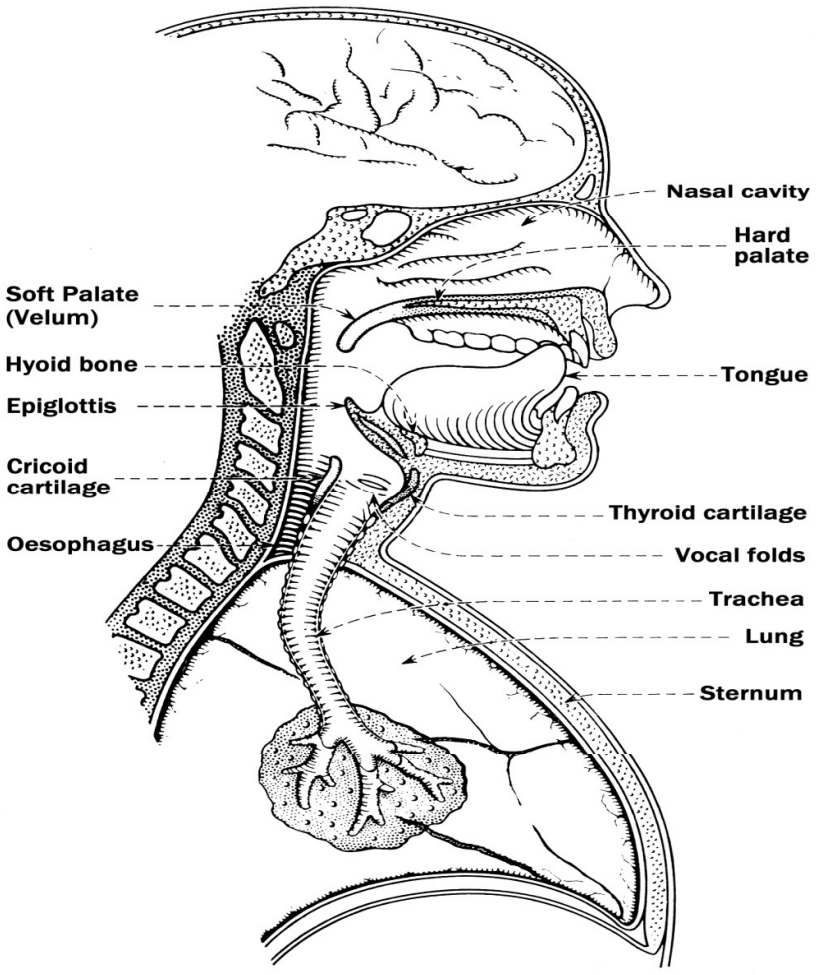



Figure 2.2. The Human Vocal System (Flanagan, 1972)

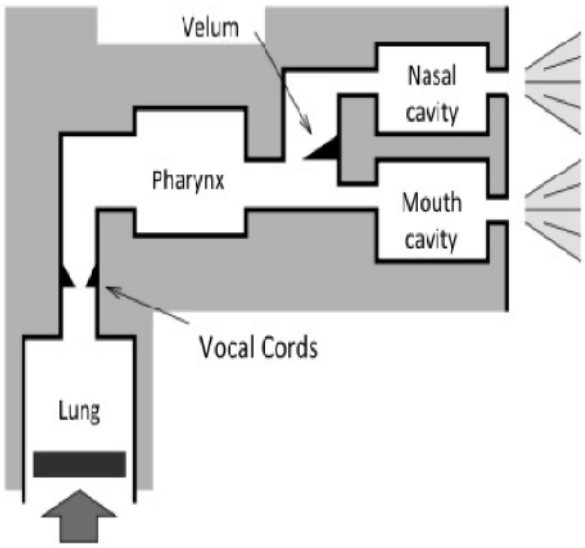



Figure 2.3. Schematic Representation of the Vocal Organ (Source: Jyothi, 2016).

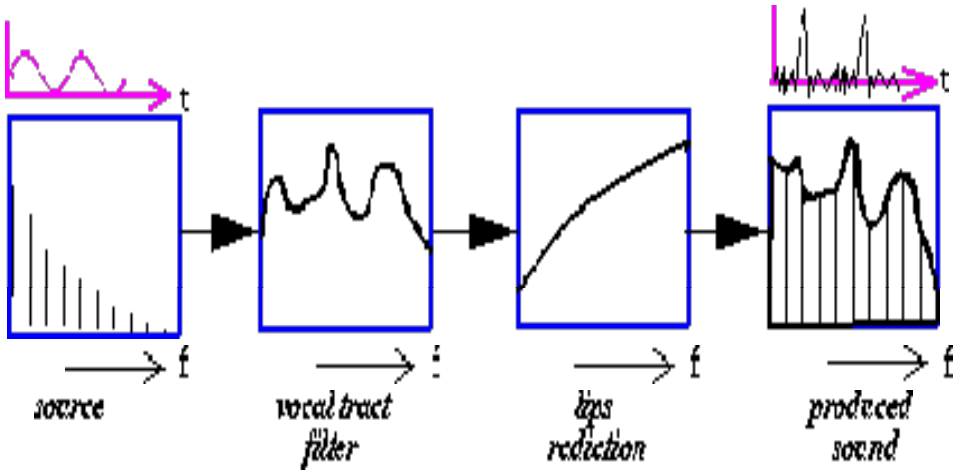



Figure 2.4. Signal waveform representation of stages in speech production. (Source: https://www2.ims.uni-stuttgart.de/EGG/page4.htm. Downloaded: April 30, 2018).

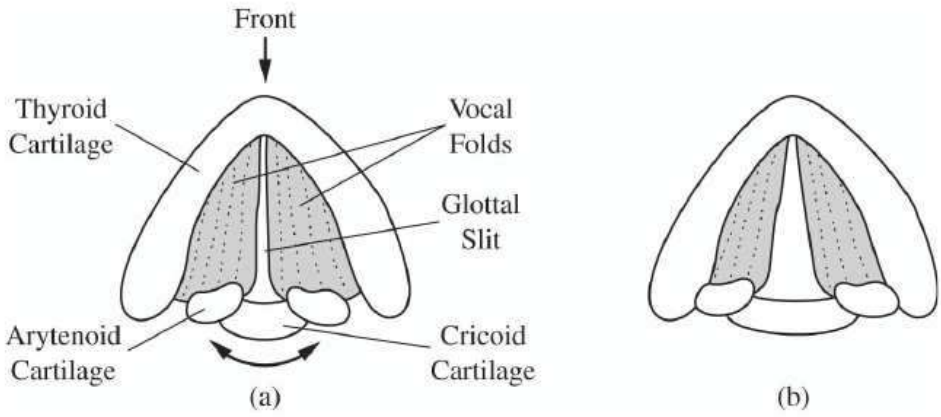



Figure 2.5 Diagram of the Larynx showing the Vocal Folds (a) Glottal Slit closed (b) Glottal Slit opened (Stevens and Weismer, 2001).

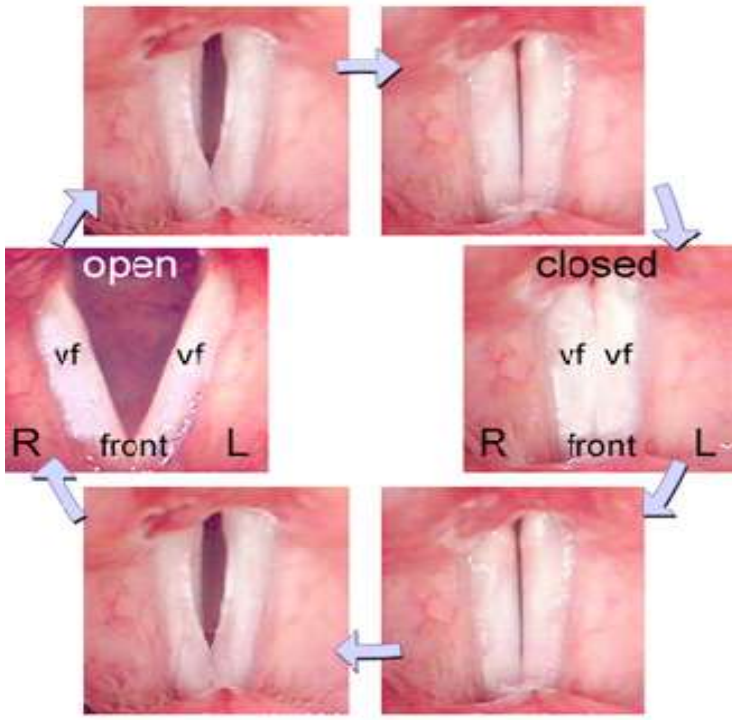



Figure 2.6. Vocal Folds Opening and Closing Positions (Source: Voiceproblem.org).

If volume velocity $u(x,t)$ and area $A(x,t)$ are used to represent the vector velocity **v** under planar assumption, then equations 2.3 and 2.4 are reduced to the following:

$$-\frac{\partial u}{\partial x} = \frac{1}{\rho x^2}\frac{\partial (pA)}{\partial t} + \frac{\partial A}{\partial t} \tag{2.5}$$

$$-\frac{\partial p}{\partial t} = \rho \frac{\partial \left(\frac{u}{A}\right)}{\partial t} \tag{2.6}$$

where$x$ is the distance from the glottis to the lips, $A$ is the cross-sectional area of the vocal tract, $u$ is the velocity of the volume of air through the glottis, and other parameters ($t$, $p$, and $\rho$) are as defined for (2.3) and (2.4).

Solutions can be obtained to these equations numerically if $A(x,t)$ and boundary conditions at the lips and at the source (the glottis) are specified.

In the upper respiratory tract are resonators known as the vocal tract, whichconsists of:the pharyngeal, nasal and oral cavities. They are known as the articulators and consist of the velum, tongue, lower jaw, and lips. These organs of articulation resonate and modulate the voice source to produce intelligible speeches. They also generate some additional sounds for consonants.

Speech is a continuous and very dynamic time-frequency signal formed by the rapid changes occurring in the movements of the vocal folds and in the vocal tract organs (Deng and O'Shaughnessy, 2003). The level of parameters such as loudness, pitch and quality of the voice, and also the generation of prosodic patterns of speech, are known to be determined by the main processes of phonation and articulation, which summarize the processes of respiration, voice formation and articulation (Honda, 2008).

The voice energy produced by the source through air streams exhaled by the lung is linearly filtered by the vocal folds, as showed in Figure 2.7, to produce the quasi-periodic sound. The resulting sound is emitted as speech into the air by radiation.This process is explained with the use of the Gunnar Fant source – filter model shown in Figure 2.8(Muñoz-Mulas et al, 2013). In this figure the excitation

signal,e(n), required for speech production is generated by phonation (voicing) or turbulent excitation or white noise (voiceless). Organs for articulation produce a pre-radiated speech signal $s_a(n)$,while the speech signal, $s_r(n)$, is produced bythe radiation or lip model.


Figure 2.7. The Vocal Tract as a Filter.

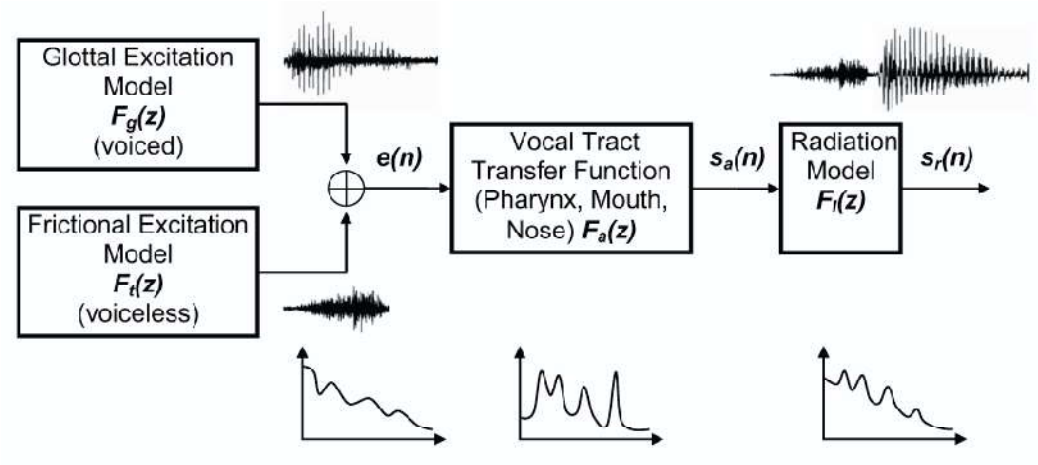



Figure 2.8. Block diagram of the Fant source-filter model (Source: Muñoz-Mulas et al, 2013).

In the study of the human voice source, different approaches of the technique of inverse filtering are adopted in both the frequency and time domains as shown in Figure 2.9 (Selent, 2014). The essence of this technique is to understand the voice source signal by simply employing an inverse transfer function of the filtering operations of the vocal tract. The inverse filter reverses the initial filtering of the source signal.

The first part of Figure 2.9 shows the voice production from the source (lung), to the filtering effect of the vocal folds to the stage of articulation and speech output. Second part shows the inversion of the speech production process. The relevant transfer functions in frequency and time modes are shown in the second and third parts of the Figure.

## 2.6 Speech Characterization

Approaches at characterizing speech signals are via their time and frequency representation (Orovic and Stankovic, 2010; Rabiner and Juang, 1993):

### 2.6.1 Time Representation

This is a display of changes in the shape and regularity of speech signal waveforms. With time characterization known as temporal structuring of speech, events in speech are categorised with respect to the state of the vocal cords. A three state representation used are: Silence (S) –no speech is produced, Unvoiced (U) – vocal cords not vibrating and speech waveform produced is aperiodic or random in nature, and Voiced (V) – vocal cords are tensed and so vibrate periodically when air flows from the lung with speech waveform produced being quasi-periodic. This classification process is based on the use of a Voice or Speech Activity Detector (VAD or SAD) (Falk et al, 2005).

The VAD or SAD is an algorithm written to identify each speech frame as active or inactive (silenced), while the active frames are labeled voiced or unvoiced. Chogule and Chavan (2014) made us to know that separating speech signal from the

non-speech part such as silence and noise is a known fundamental problem in many speech processing systems. In speech coding and transmission, the unvoiced and silence frames are removed, so that only the voiced frames are available for computation or transmission. This helped in significantly reducing computation time

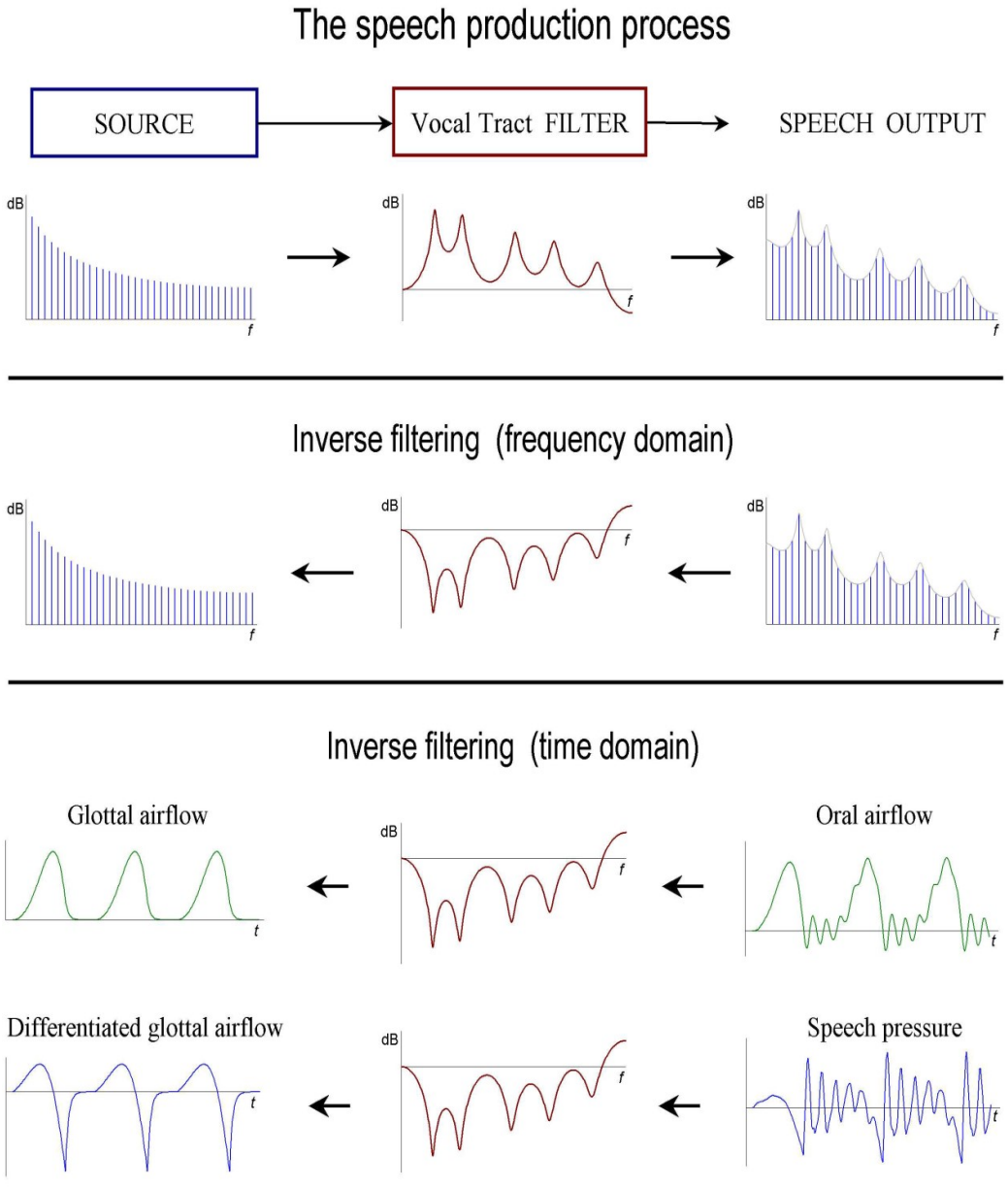



Figure 2.9. Inverse Filtering in Frequency and Time Domains (Selent, 2014)

in speech processing systems and bandwidth requirement for transmitted speech on telecommunication networks.

Noise reduction techniques used in speech processing and speech communication systems require that noise statistics be estimated, have adopted voice activity detection for this purpose (Ramírez et al, 2007). This made VAD to also be applied in systems for speech enhancement and robust speaker recognition.

Different methods of designing and building VAD algorithms generally make use of speech features such as periodicity measure, energy thresholds, pitch detection, zero-crossing rate, spectrum analysis (Fourier coefficients), and so on. Modern VADs while making use of these features based their speech or non-speech discriminations on heuristics or on either supervised or unsupervised learning statistical model approaches such as Gaussian classifiers or Laplacian models (Ramírez et al, 2007; Ying et al, 2011; Kola et al, 2011).

Three VAD algorithm methods reported by Ramírez et al (2007) are: long-term spectral divergence, multiple observation likelihood ratio tests, and order statistics filters. In their work, (Kola et al, 2011) compared four major types of VAD algorithms – the Ying VAD algorithm (Ying et al, 2011), the Sohn VAD algorithm (Sohn et al, 1999), the ITU G.729B VAD algorithm (ITU Rec G.729, 1996), and the two implementation of the ETSI AMR VAD algorithm – AMR1 and AMR2 (ETSI EN 301 708 Rec, 1999).

Their work was to find out which of these algorithms performances best and most consistent when tested on the same set of speech utterances, types of noise as non-speech signals and SNRs. Of these, the Ying algorithm was most outstanding in all noise types and SNRs.

### 2.6.2 Spectral Representation

This is spectral structuring of speech with the use of sound spectrogram for displaying multi-frequency bands speech intensity in 3–dimensional representationover time. In the period covering unvoiced speech, there is mainly high-

frequency energy in the spectrogram, while during silence periods there is no spectral activity.

Both temporal and spectral structures of speech possess non-stationary properties through which information conveyed by speech signals could be described. But in describing the temporal evolution of speech parameters particularly in very low bit rate speech encoding used for speech transmission, speech synthesis and speech recognition, stationary signal processing tools like Discrete Fourier Transform (DFT) and Linear Predictive Coding (LPC), and parametric and non-parametric techniques are used (Ahlbom et al, 1987).

Temporal decomposition of speech first proposed by Atal (1983) as a method for carrying out major reduction in information for representing the spectral characteristics of speech has witnessed the use of techniques like the Singular Value Decomposition (SVD) approach and the Preferred Iterative Approach (PIA) (Bailly et al, 1989). Ritz et al (2000) represented speech excitations by utilizing Characteristic Waveform (CW) that was extracted at a constant rate in temporal decomposition of speech.

## 2.7 Psychoacoustic Features of speech

Speech features are the speaker-specific information found in a speech signal. These are: voice source features, short-term spectral features, spectral-temporal features, prosodic features (that is, syllable stress, intonation patterns, speaking rate and rhythm of speech, in linguistics), and high-level features (conversation-level features, e.g. characteristics vocabulary of speakers) (Yankayis, 1991).

### 2.7.1 Short-Term Spectral Features

The spectral envelope of a speech signal displays peaks or vocal tract resonances of the signal known as the formants, which are the dominant frequency components of the speech signal shown in Figure 2.10. Formants are in inverse proportion to the length of vocal tract and they stand for the identity of the speech sound. On the speech spectrogram, features (formant) derivable from the vocal tract characteristics are shown in Figure 2.11.

Formants, as short-term features, are extensively utilized in speech recognition applications based on the fact that for a given sound, different individual speakers (male, female, adult or child, or sour-throated individuals, and so on) will have

different spectral shapes. For different speakers and different sexes, the characteristics of the vocal tract, the location and sizes of the formants are never the same.

### 2.7.2 Voice Source Features

These are glottal pulse shape and fundamental frequency of the speech signal, and are specified by the rate and level of oscillation of the voice folds. This oscillation is determined by factors like the mass and length of the voice folds and the tension exerted on it by the muscle. Glottal pulses passing through the vocal tract are filtered by it, making it difficult to directly measure the voice source features from the speech signals.

A way out is to carry out an inverse filtering, S(z), of the speech signal, discussed in section 2.5, and depicted by:

$$S(z) = E(z)\frac{1}{H(z)} \tag{2.7}$$

where H(z) represents the transfer function of the vocal tract, while E(z) is the signal from the speech source, the lung.

In estimating the characteristics of the vocal tract filter the Linear Predictive Coefficient (LPC) model can be used. The voice source features depend more on the pitch of the sound produced by the vocal folds (Yankayis, 1991).

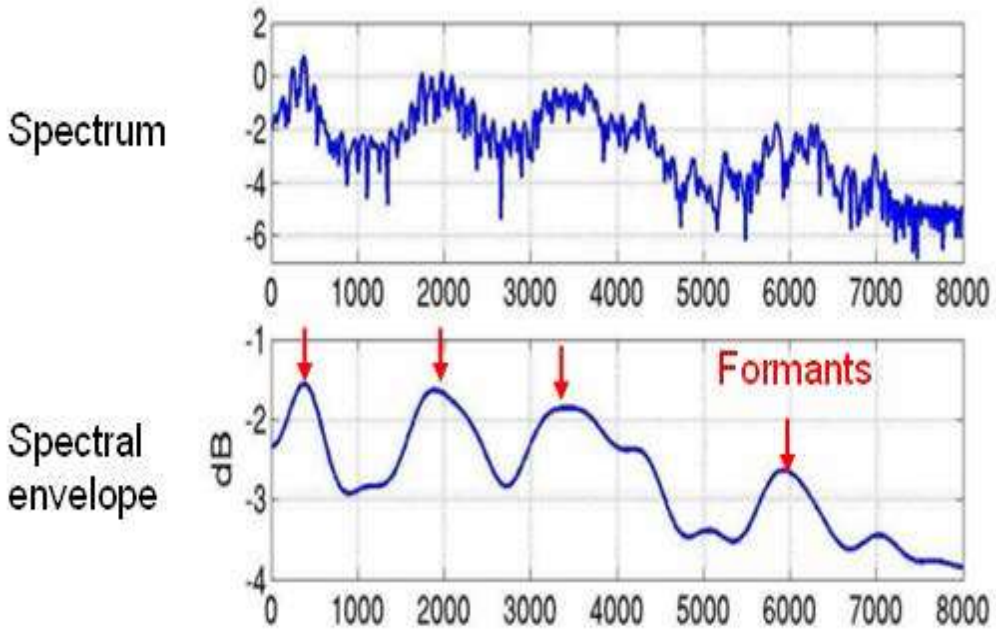



Figure 2.10. Spectral Peaks (Formants) of a Speech Signal (Source: Yankayis, 1991).

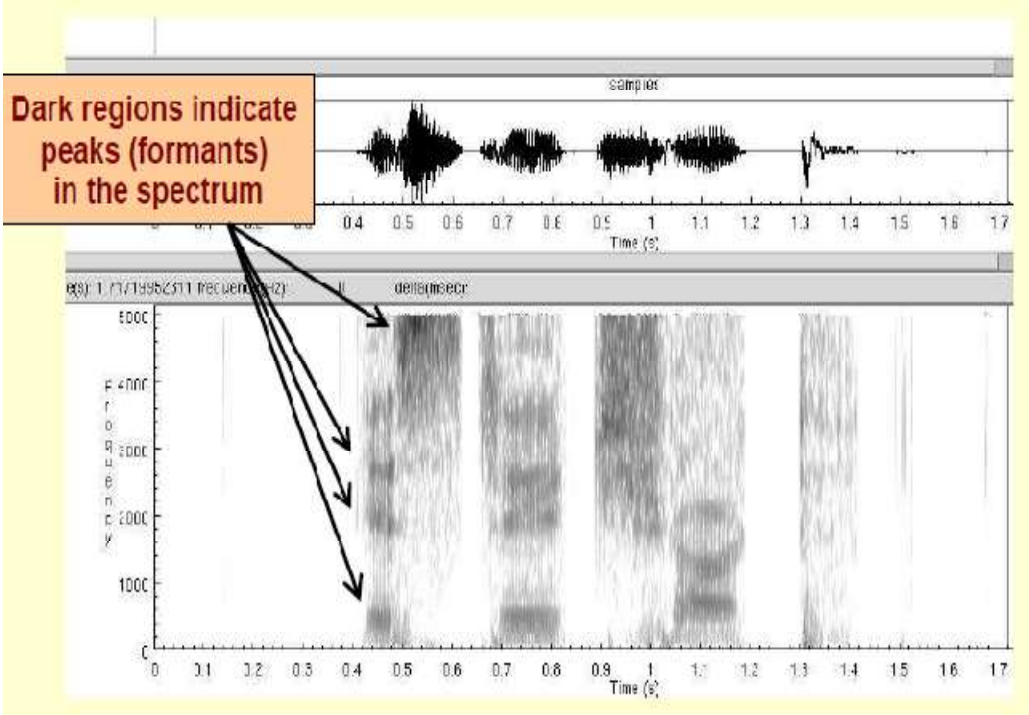



Figure 2.11. Formant Representation on Speech Spectrogram (Yankayis, 1991).

## 2.8 The Human Auditory System

Psychoacoustics helps in understanding the response effect of the hearing system on sounds entering into it and also sheds light on the anatomy of the human auditory system. In it is the path for incoming sounds up to where it is conveyed by nerve fibers to the brain. The quality of voice signal received by the end-user of telecommunication services is rated as perceived by the human auditory system. Computational models are built and used for quantitative simulations that help to describe the functionalities of the auditory system and its responses to and perception of transmitted voice signals.

The human hearing system known as the peripheral auditory system is shown in Figure 2.12, and consist of three major parts: outer, middle and inner ear, is hereby described: (Pulkki and Karjalainen, 2015; Kollmeier, 2008; Rabiner and Juang, 1993).

The auditory system feeds sound waves through various processes of impression and transduction into the neural system for perception and interpretation by the brain. Starting at the outer ear, there is the pinna (or auricle), the ear canal and the eardrum. Sound waves enter the pinna (the external flap) and go through the auditory canal (a hollow of about 2.0 to 2.8cm long), to impinge on the eardrum (also known as tympanic membrane). The eardrum is a light, thin, highly elastic structure which closes physical access to external materials and air.

The resulting mechanical vibration of the eardrum by the acoustic air pressure entering the ear is transmitted to the ossicles in the middle ear. The ossicles are three small bones, namely: malleus, incus and stapes, commonly known as hammer, anvil and stirrup respectively. The ossicles act as impedance transformers, maximizing energy transmitted into the inner ear and eliminating reflection of waves at the boundary of gas and liquid mediums in the ear.

The malleus acting as a hammer is fixed unto the eardrum on one side and strikes the incus as an anvil, which is joined to the stapes or stirrup. The stapes is attached to an oval hole (or window) on the inner ear's cochlea. Vibration of the

stapes by the incus causes a stirrup in the fluid (perilymph) contained in the cochlea. This leads to displacements in this fluid and variations of the pressure within the cochlea. Acoustic vibrations entering the ear are therefore transmitted by the tympanic membrane via the ossicles as mechanical movements to the cochlea in the inner ear.

Cochlea is the major component of the inner ear. It has a snail-like structure in form of a spiral-coiled tube of approximately 2.75 turns. It converts the mechanical

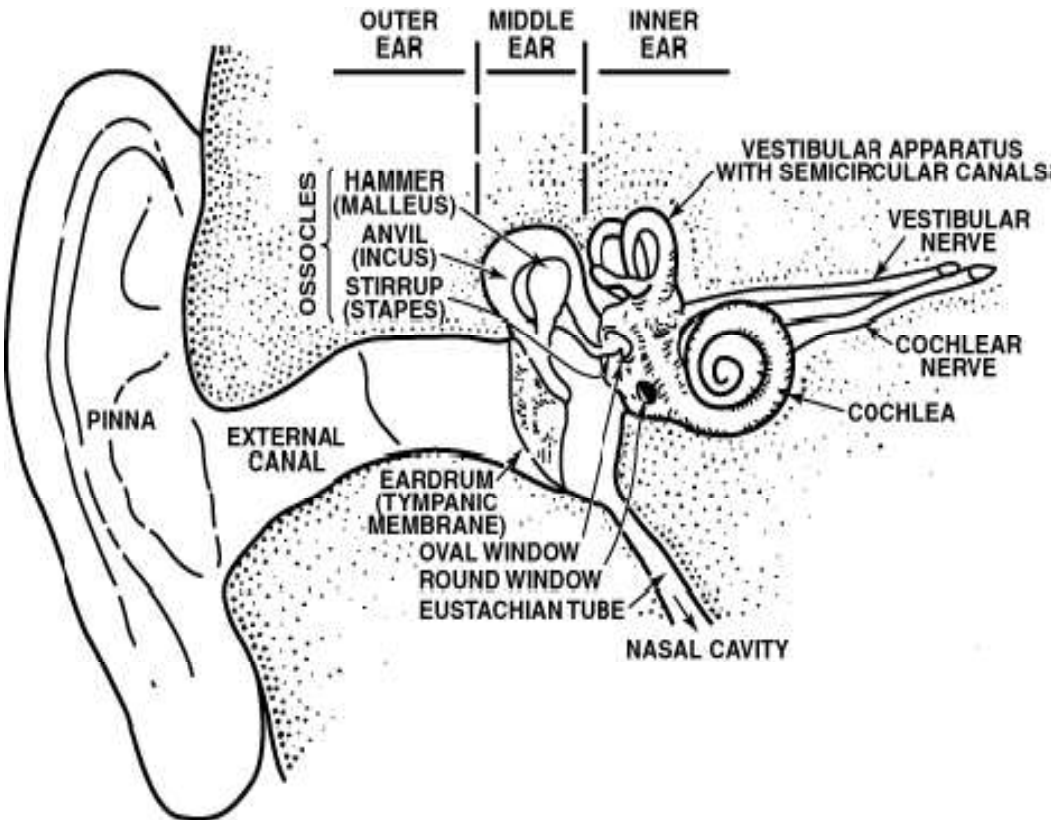



Figure 2.12. Anatomy of the Human Auditory System. (Source: Rabiner and Juang, 1993).

vibrations from the middle ear into nerve firings for processing and interpretation by the brain. The end of the cochlea, connected to the middle ear stapes, is the "base" and has the oval and round windows. The other end of the cochlea is the "apex." The width at the base is about 0.04 mm and about 0.50 mm at the apex (Lin and Abdulla, 2015). Shown in Figure 2.13 is the cross-section of the cochlea when stretched out and sliced through (Howard and Angus, 2009).

As the stapes move in and out at the oval window, the incompressible fluid in the cochlea(the perilymph)is pushed and relaxed, making the round window on its other side to deflect in and out correspondingly. Consequently, travelling waves are created inside the cochlea which cause the basilar membrane to move, and the bundle of inner hair cells (stereocilia) in the organ of Corti of the cochlea to be displaced. These phenomena in the cochlea are responsible for variations in the input impedance of the inner ear with the level of vibration and frequency of sound to which the ear is subjected. The acoustic vibration through the basilar membrane is propagated as traveling waves through it, with a point of maximum amplitude of the vibration induced by the frequency-dependent travelling waves (Volk, 2016).

Distributed along the Basilar Membrane (BM) are hair cells, which are bent by movement of BM, and consequently triggers nerve firings to the brain. The nerve firings are a kind of electrical signals transmitted to the brain through a spiral bundle of auditory nerves connected to the hair cells. The neural excitations travel to the brain through what is known as auditory pathway (Kollmeier, 2008; James et al, 2018).

## 2.9 Auditory Filter Bank and Critical Bandwidth

BM serves to analyse sounds entering the ear in the frequency domain. Its typical shape shown in Figure 2.14 has different frequency response to stimulations by sound signals at different points along its length. The BM vibrates differently to sounds of frequencies ranging from lowest to highest on the sound frequency scale (20 to 20,000 Hz) audible to the human ear (Cote, 2011).

BM's response to different sound frequencies is determined by its mechanical properties, as they progressively vary from the base to the apex. The BM is relatively thin, narrow and stiff at the base, making the base to respond best to high frequencies. The BM is thicker, wider and much less stiff at the apex, making the apex to respond

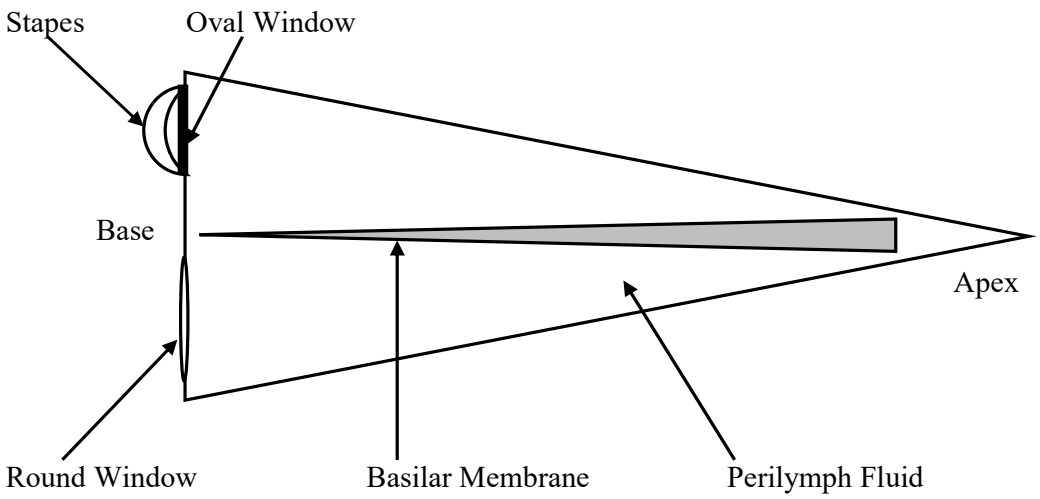



Figure 2.13. Illustration of the Cochlea Vertical Cross-Section. (Source: Howard and Angus, 2009).

best to low frequencies. Each point on the BM responds with greatest displacement to a certain frequency, known as the characteristic frequency (CF), and responds less and less as the frequency is moved farther from CF (Moore, 2003). With this, we say that BM is tuned to such particular frequency. For example, a 1,000 Hz tone would most strongly stimulate that part of BM which has a characteristic frequency of 1,000 Hz.

For complex sounds with multiple frequency components, the overall response of BM is a sum of the responses for each of the components. This forms the basis of what is known as the "place" analysis of sound by the auditory system (Howard and Angus, 2009).

According to "place" theory, basilar membrane acts as a tuned resonator that extracts spectral representations of incoming sounds (Oxenham, 2008). The point of resonance shown in Figure 2.14c linearly relates to frequency, with the characteristic frequency gradually decreasing as one move from the base to the apex of the cochlea. The frequency – to – place mapping is known as tonotopic mapping, resulting in a neural spectrogram transmitted to the brain (Oxenham. 2008; Cote, 2011). The tonotopic map of the human cochlea in Figure 2.15 shows the representation of real-life sounds with complex harmonic tones in the auditory system.

Along the cochlea's basilar membrane each point responds to certain frequency range similar to the bandwidth of a band-pass filter. This implies that BM's whole length can be represented with a bank of bandpass filters, rather known as auditory filters, with overlapping frequency resolution and bandwidth with their center frequencies covering the whole audible range. The values of the filters' bandwidths known as Critical Bandwidth (CB), characterize the center frequencies of the filters. These non-uniform frequency bands are shown in Figure 2.16.

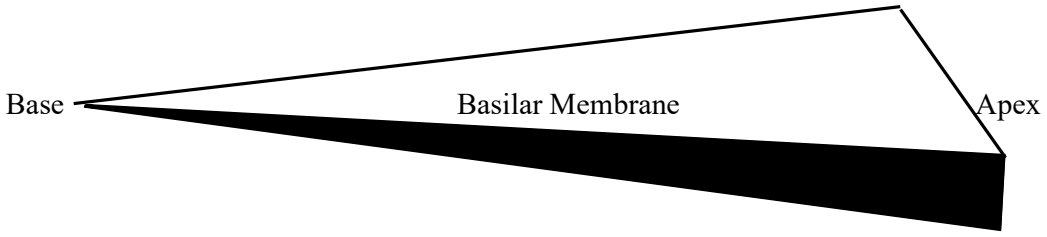
The bandwidth varies with the value of the center frequency, that is, it increases with increasing center frequency (Wang et al, 1991). Using Patterson's method of expressing this relationship, (Moore, 1987) noted that over the frequency range 100 to 6,500 Hz this bandwidth can be expressed by:

$$\mathrm{BW} = 6.23 f_c^{\,2} + 93.39 f_c + 28.52 \qquad (2.8)$$

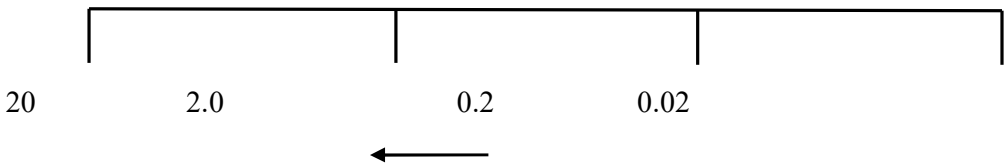
where $f_c$ is the center frequency of each band.

The bandwidth over the whole human hearing range was given empirically by (Rabiner and Shaffer, 2007):

$$\Delta f_c = 2.5 + 75[1 + 1.4(f_c / 1000)^2]^{0.69} \qquad (2.9)$$

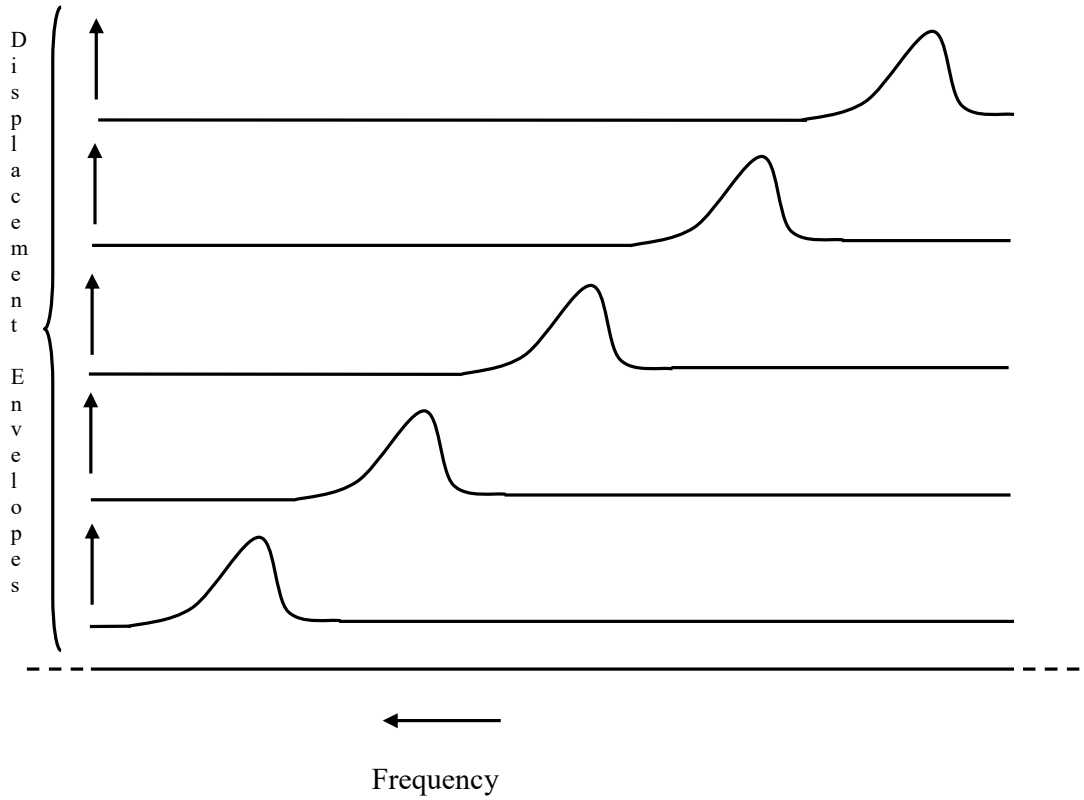



(a) Idealized shape of the basilar membrane




Frequency (kHz) - logarithmic

(b) "Place" frequency response of the basilar membrane



(c) Points of maximum displacement with respect to frequency

Figure 2.14. Idealized Shape and "Place" Frequency Response of the Basilar Membrane. (Source: Oxenham, 2008).

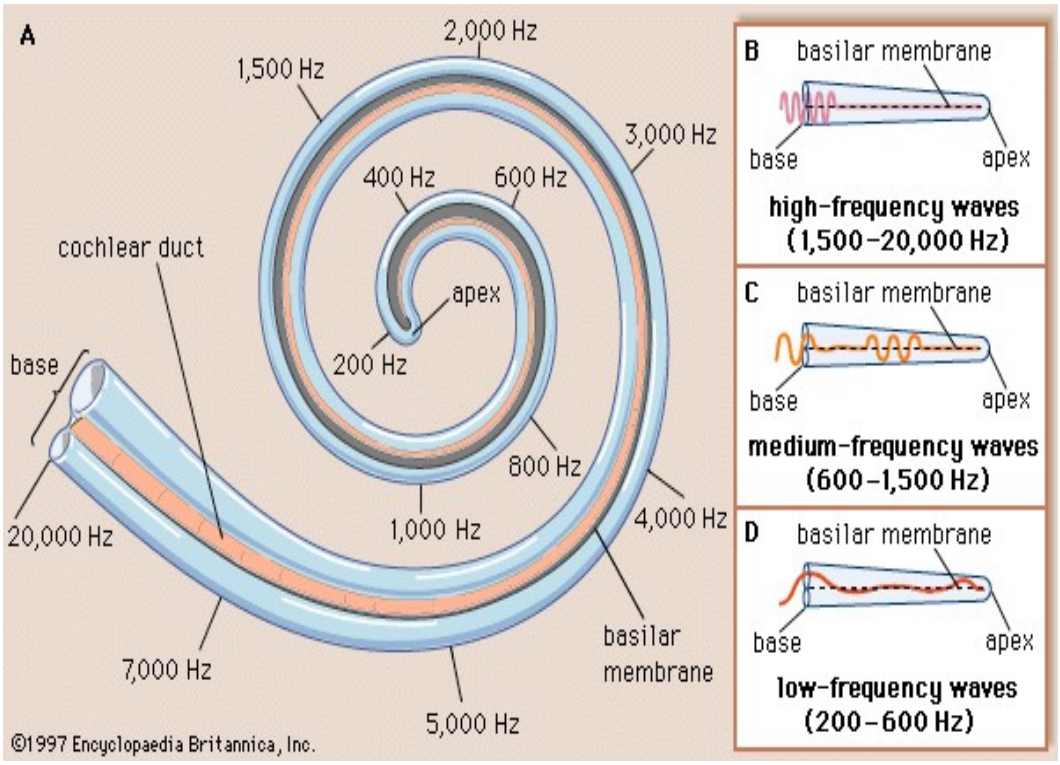



Figure 2.15. Tonotopic Map of the Human Cochlea.(Source: Oxenham. 2008).

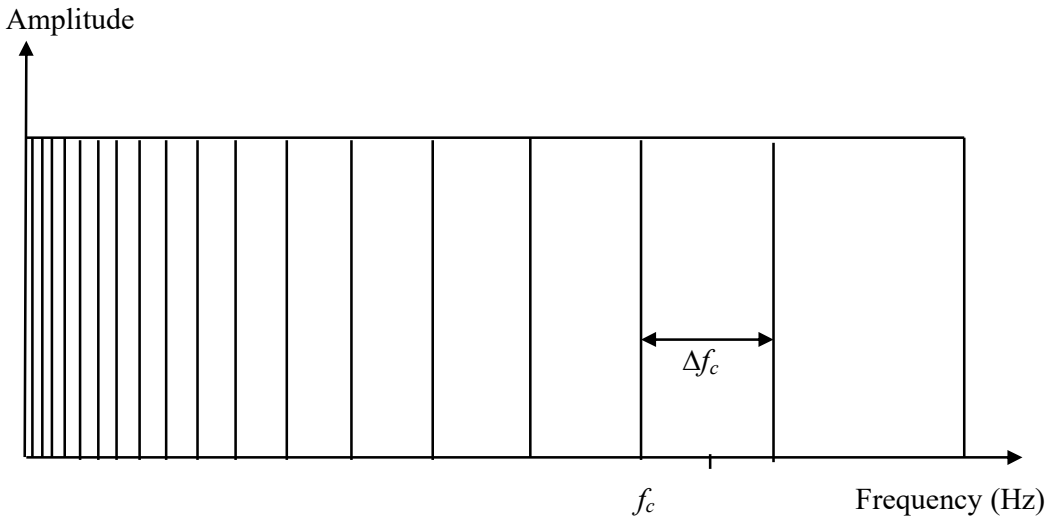



Figure 2.16. Critical Bandwidths of the Human Auditory System.

In investigating the characteristics of the auditory filters, Fletcher was among the pioneers of the application of the concept of auditory-filter for masking sinusoidal tones by broad-band noise (Moore 1987). With the use of a rectangular shape having a flat top and vertical edges, he approximated the auditory filter's response which enabled the prediction of thresholds. This means that all frequencies within the flat top or pass-band would be equally passed, and others rejected.

With the use of masking and other psychoacoustic techniques, the auditory filters' bandwidths have been estimated and realized with cascaded Infinite Impulse Response (IIR) filters in the design of a psychoacoustic model for improving audio coding as shown in Figure 2.17 (Baumgarte, 2002). Each Low-Pass Filter (LPF) is connected to a High-Pass Filter (HPF) with a cutoff frequency equal to that of the LPF cascade segment between the input of the filter-bank and the input of the HPF of the next section. This implies each HPF's output has a band-pass characteristic with respect to the input signal of the filter-bank.

In the multi-channel analysis of the cochlea it is noted that the channels overlap, and the bandwidth of each channel is equivalent to one critical band or 1 Bark. The frequency range (20 – 20,000 Hz) to which the human ear is sensitive is equivalent to CB rates within the range 0 – 24 Barks by the equation developed by (Zwicker et al, 1957) and given by (Cote, 2011):

$$z_B = 13 \arctan(0.76\, f) + 3.5 \arctan\left[\left(\frac{f}{7.5}\right)^2\right] \tag{2.10}$$

where, $f$ is frequency in kHz and $z_B$ is CB rate in Bark.

In another development, the critical-bands are represented in Equivalent Rectangular Bandwidth (ERB) of the range 0 – 40 ERB and given by Cote (2011):

$$z_{ERB} = 21.4 \log_{10}(4.38\, f + 1) \tag{2.11}$$

where, $f$ is frequency in kHz and $z_{ERB}$ is CB rate in ERB.

Hence, for computational modeling purpose, one to four channels are specified per Bark. Since the whole audio range consists of 24 Bark, it implies that a total of 24– 96 channels are required for computational models, and a 0.5 Bark spacing leads to 48 channels for practicable computations (Karjalainen, 1987).

Some theories and techniques that have been associated with the auditory filterbank include the auditory frequency-scale warping, and constant-Q filterbanks built on wavelet transform and known as auditory wavelet filter bank. Others are different psychoacoustic testing, resulting in the auditory filter frequency response

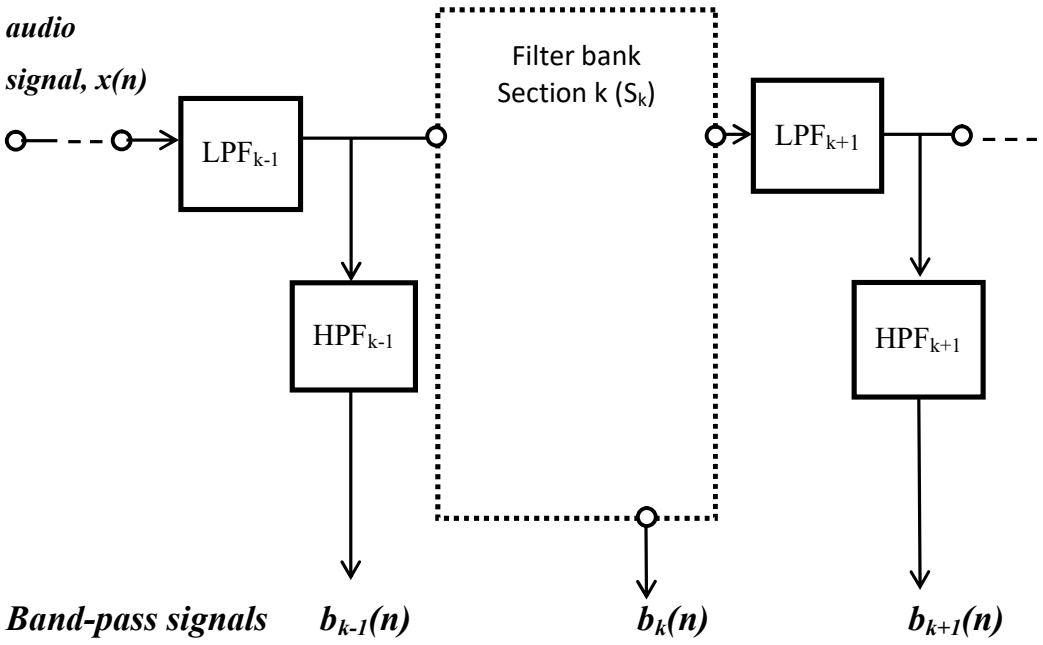



Figure 2.17 Block Diagram of a Cochlea Filter-Bank Structure (Source Baumgarte, 2002).

being approximated with the use of Gaussian functions, a rounded exponential, and the gammatone and gammatone-chirp filterbanks (Irino and Patterson, 2006; Smith and Abel, 1999; Irino and Unoki, 1998).

## 2.10 Human Hearing Range

The human hearing range covers three decades of frequency from 20 Hz to 20 kHz for normal ear. With the speed of sound at 343 m/s in air, the wavelength of audible sound varies from 17 to 1.7 m. The audibility curve in Figure 2.18 shows the threshold of hearing and changes on this curve shows that humans are most sensitive to sounds between 2,000 and 4,000 Hz. The threshold of hearing which is the minimum Sound Pressure Level (SPL) of a pure sinusoidal tone detectablein the absence of any other sound, is given in dB by (Zwicker and Fastl, 2007):

$$SPL = 20\ log_{10}(P/P_o)\ \text{dB} \tag{2.12}$$

where,$P$ is sound pressure and $P_o$ is reference sound pressure normalised to 20 $\mu Pa$.

The other curves in Figure 2.18 are some of the curves discussed in equal-loudness curves chart in sub-section 2.11.1 and the threshold of feeling where pain is felt in the auditory system with increased sound pressure level (SPL).

## 2.11 Perceptual Psychoacoustic Properties of Sound

General psychoacoustic properties of sound that help in describing and analysing its perception and perceptual quality by the human auditory system are known as psychophysical metrics or parameters of sound (Zwicker and Fastl, 2007). They include:

1. Loudness
2. Sharpness
3. Pitch
4. Roughness
5. Timbre
6. Fluctuation Strength

Perceptual factors that affect the way speech signals are perceived by listeners identified by (Moller, 2000) are as follows:

1. loudness – results in the concept of loudness rating,
2. articulation,
3. how effects of bandwidth and linear frequency distortion are perceived,

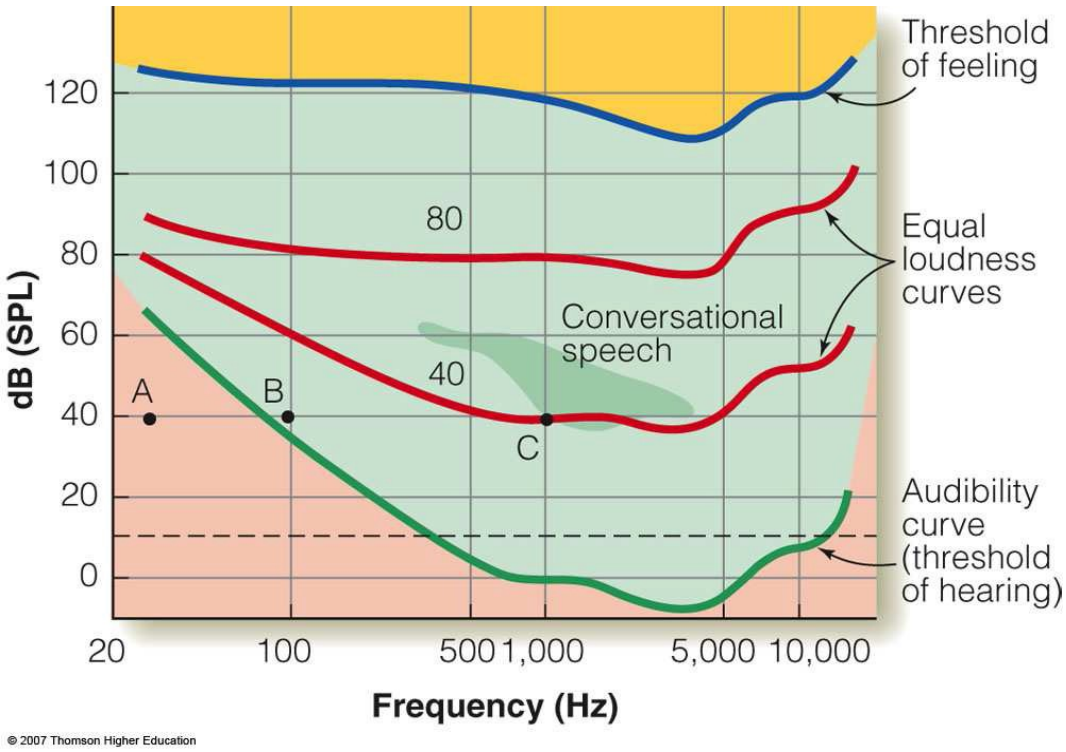



Figure 2.18. Threshold of Hearing. (Source: Zwicker and Fastl, 2007).

4. how one's own voice is perceived (sidetone),
5. how echo is perceived,
6. how circuit noise is perceived (continuous, impulsive, bursts),
7. effects of environmental noise and binaural hearing, and
8. effects of delay.

### 2.11.1 Loudness

Loudness is the magnitude or intensity of the resultant sensation of sound of any quality or structure that impinges on the auditory system.Loudness level is a measure of comparison made to characterize the loudness sensation of any sound. Introduced in the 1920's by Barkhausen, after who the critical-band rate (Bark) was created, loudness level of a sound was equated to the Sound Pressure Level (SPL) of a 1 kHz plane wave tone that is as loud as the sound, with rated in the unit "phon" (Zwicker and Fastl, 2007).

Figure 2.19 shows the "Equal-loudness contours," consisting of loudness levels of pure tones of different frequencies plotted within the hearing area and the lines connecting points of equal loudness within each area. The curves go through the SPL at 1 kHz having equal value (in dB) as the parameters of the curve (in phon). Examples are the 60 phon going through 60 dB point at 1 kHz and the threshold of hearing in quiet (dashed curve) seen to correspond to 3 dB at 1 kHz.

In an attempt to describe perception of sound, the equal-loudness contours provide a scale which enables us to compare the loudness of different sounds. The frequency range over which a complex sound like the human speech extends helps to determine its loudness, provided the total intensity of the sound is fixed (Moore, 1987). That is, we hold sound intensity as a constant while we vary the bandwidth, and then we vary sound intensity while we make bandwidth a constant at a particular point. These lead to varying the loudness in both cases and the two cases would sound equally loud. It also implies that two sounds of equal intensity but different frequencies will not have the same loudness because of the varying sensitivity of auditory system to frequency.

In measuring loudness sensation, sound intensity level is used relative to a sound of frequency 1 kHz and level of 40 dB proposed to give the reference loudness sensation as 1 Sone. In evaluating loudness, the doubling or halving ratio is used. By implication, the SPL of a sound has to increase by 10 dB for the loudness sensation to

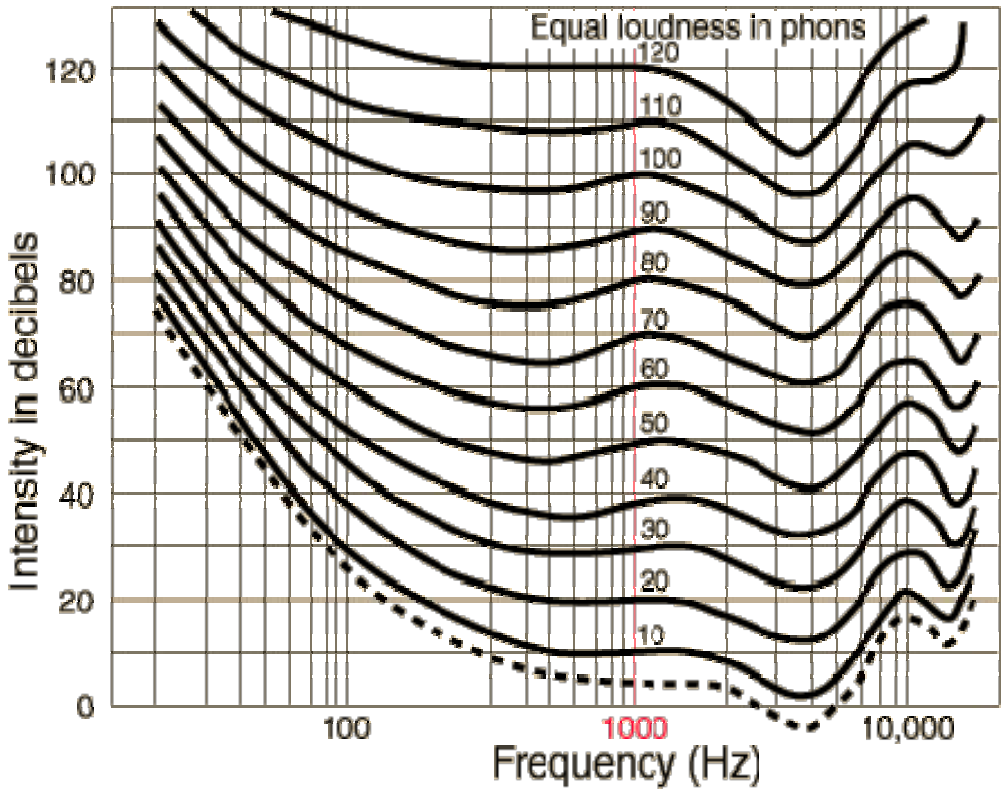



Figure 2.19. The Equal-loudness Contours Curves. (Source: Zwicker and Fastl, 2007).

double, that is, an increase of 2 Sones, in a form of power law. The loudness function is plotted by halving and doubling of loudness over the whole loudness range. Loudness function given for 1 kHz can also be used to plot for otherfrequencies with the aid of the equal-loudness contours.

Subjective loudness, $S_L$, in 'sone' is related to sound intensity by the Steven's power law given by Moore (1997) and Bernasconi and Seri (2016):

$$S_L = C\,(W^{0.3}) \tag{2.13}$$

where, $C$ is the constant of proportionality and $W$ is the intensity of sound.

Models for predicting and measuring the loudness of speech and sound both subjectively and objectively have been proposed over the years, twelve of which were reviewed by Skovenborg and Nielsen (2004). The Zwicker loudness model standardized in ISO 523 in 1975 and reviewed into the ISO 523B in 2002 provides quantitative description of calculating the loudness of steady-state or stationary tones. It can be used to precisely determine specific loudness and loudness pattern of any sound (Boilot and Harris, 2004).

The model developed by Glasberg and Moore (2002) based on excitation loudness pattern,calculated the loudness of time-varying sound using the waveform as input was formally standardized in ANSI S3.4. (ISO 532-1:2017). Moore and Glasberg (1996)made efforts at quantifying the loudness of speech with respect to characteristic estimation of glottal excitation by speech signals. This was based onfactors of the vocal tract and the articulators, resulting in variabilityofloudness of speech.

To estimate the loudness of different non-stationary sounds, approximation is done to it through arithmetic averaging of the SPL in octave bands.Arithmetic averaging of SPL in octave bands for 63 Hz to 4 kHz, that is$L_{\mathrm{m},1/1}$(63 – 4k), is very similar with Loudness Level measured with Zwicker model, $LL(Z)$, defined in ISO 532B and defined by Simpson et al (2013) as:

Loudness Level,$LL(Z) = 40 + 10\log_2 N$ (in phon) (2.14)

where,$N$ is the total loudness (in Sone).

A method for obtaining instantaneous loudness of time-varying sound along with the auditory excitation patternswas described by Chen and Hu (2012). They affirmed loudness as being dependent on sound duration, and that for a sound with fixed intensity, the loudness increase as duration increaseswithin 100 to 200 ms and when the duration of less than 100 ms is doubled, the loudness increase by 3 phons. Increase in the loudness of sound with respect to duration is known as Temporal Integration of Loudness (TIL) and for sound durations exceeding 200 ms, loudness do not increase any more (Glasberg and Moore, 2002; Zwicker and Fastl, 2007; Pulkki and Karjalainen, 2015).

Mindful of the complexity of algorithms/models for estimating loudness, Krishnamoorthi et al(2008) developed a low-complexity model useful for steady and time-varying sounds utilising the Glasberg and Moore's model. His work was carried out by computing a fast estimate of the excitation pattern through selecting the most relevant frequency component locations in a uniform manner. Rennies et al(2010) made a comparison of various models to calculate the loudness of time-varying sound, and considered principal model differences and their limitations.

Critical bandwidth isan important consideration in determining loudness and the excitation level versus critical bandwidth rate pattern of the auditory system which are used as basis for obtaining the loudness for complex sound such as speech or music (Zwicker and Fastl, 2007). Total loudness was therefore givenas an aggregation of specific loudness, $N'$, over the critical-band rate, as follows:

$$N = \int_0^{24\ Bark} N'\, dz \qquad (2.15)$$

where specific loudness is obtained from:

$$N' = 0.08 \left(\frac{E_{TQ}}{E_0}\right)^{0.23} \left[\left(0.5 + 0.5\frac{E_{TQ}}{E_0}\right)^{0.23} - 1\right] \frac{\text{sone}_\text{G}}{\text{Bark}} \qquad (2.16)$$

where,$E_{TQ}$is excitation at threshold in quiet, and $E_0$ is excitation that corresponds to the reference intensity($I_0 = 10^{-12}$ W/m$^2$).

For sounds of continuous spectrum, for example speech transmitted over telecommunication networks, critical bands notation are potentially useful in determining loudness level (Moller, 2000). For the purpose of calculating loudness level, Moller made use of Figure 2.20to prove that speech spectrum and the threshold of hearing are related. Curve (a) is the speaker's speech spectrum at his mouth, curve (b) is the spectral density of the speech reaching the listener's ear, curve (c) is the hearing threshold for sounds of continuous spectrum, and curve (d) is the hearing

threshold masked by noise. The effects of a frequency dependent loss,$L_{ME}$, introduced by the transmission channel, include reduction of the particular sound's spectral density and the loudness of perceived sound in the absence of noise, which is the difference between (b) and (c) is seen as a function of $Z_{L,}$.

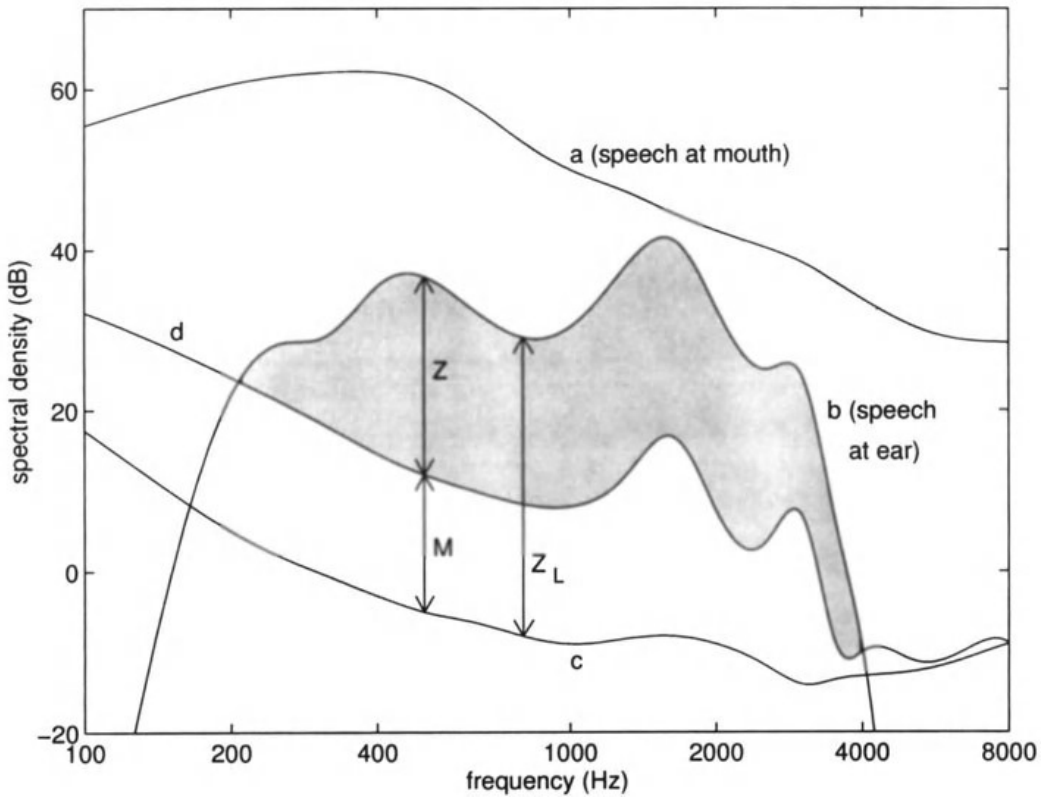



Figure 2.20. Curves of Speech Spectrum and Threshold of Hearing. (Moller, 2000).

### 2.11.2 Sharpness

Sharpness is defined as a quantity of the high frequency components of a given sound linked to the spectra characteristics of the sound, such that the greater the high frequency content of a sound the sharper the sound becomes. It is a measure of tone colour, and measured in acum (Latin word for "sharp"). 1 acum is equivalent to the sharpness of a 60 dB narrow-band noise that is one critical band wide with a center frequency of 1 kHz (Ferguson and Brewster, 2017; Zwicker and Fastl, 2007).

For narrow-band noises sharpness increase with increasing center frequency, such that at low frequency sharpness increases proportional to the critical-band rate up to 3 kHz. However, at high frequencies where critical-band rates do not increase so much, sharpnessincreases faster than the critical-band rate of the center frequency of the noise.

Sharpness, S, therefore can be obtained from the equation (Zwicker and Fastl, 2007):

$$S = 0.11 \frac{\int_0^{24\,Bark} N'\, g(z)\, z\, dz}{\int_0^{24\,Bark} N'\, dz} \text{ acum} \qquad (2.17)$$

where, the denominator is the total loudness, *N*, defined in Equation 2.15, while the numerator is the first moment of CB rate,$g(z)$ is a CB-rate dependent factor, which is critical-band rate dependent.

### 2.11.3 Pitch

Pitch is the attribute of auditory sensation responsible for ordering sounds on a musical scale and is determined by frequency, that is, the higher the frequency of a sound the higher the pitch. This dependence on frequency is mainly for pure tones, while for complex sounds like speech and music, it is determined by the period of the waveform (Ferguson and Brewster, 2017).

The mechanism underlying perception of pitch was associated with the distribution of auditory activities across nerve fibres, that is, the 'place' theory by

(Moore, 1987). Moore also related pitch to temporal theory, which portrays pitch as determinable by neural spikes in time-domain.

### 2.11.4 Timbre

Timbre was defined as tonal colouration of sound and as being the characteristic quality of sound that differentiates one voice or musical source from another one of the same loudness, pitch and duration as the other voice or music. It is a multidimensionalpsychoacoustic parameter, that depends on several other physical properties of sound, including the following (Moore, 1987; Pulkki and Karjalainen, 2015):

1. The periodicity of the sound (gives a tonal quality for repetition rates between 20 to 16 000 Hz), or the irregularity (gives a noise-like quality).
2. Whether the sound is continuous or interrupted.
3. Distribution of energy over frequency, that is the spectrum, and changes in the spectrum with time.

## 2.12 Estimating the Quality of Speech Signals

Previous works on estimating the quality of speech signals in any speech processing system or transmission networks have been based on techniques built around extracting psychoacoustic information or features from speech signals through either of two separate approaches based on the system of auditory perception and speech production mechanisms.Different speech quality assessment algorithms and models are developed around these two perspectives.

### 2.12.1 Speech Quality Assessment Models Based on Auditory Perception

Computational models that help in understanding, analyzing and simulating the human auditory system by mimicking its performance are built for assessing speech quality. These models make use of psychoacoustic information extracted from the speech waveform or its equivalent Fourier transformation (Ghitza, 1994).

Auditory models are developed and utilized in various application areas of speech processing which include speech recognition, speech analysis, speech coding, speech synthesis, speech quality measurement and technical audiology and phoniatrics (Karjalainen, 1987). These models includethe Flanagan's model, Lyon's model, Meddis Inner Hair Cell (IHC) model, Auditory Image Model (AIM), Seneff

model, Auditory Perceptual (AP) modeland Ensemble Interval Histogram (EIH) model. Some of these models were briefly reviewed in this work.

In auditory models, both original and degraded speeches are fed into the model to make comparisons of their effects on the auditory components. Specific–loudness patterns of these signals are compared as indicated by (Hauenstein, 1998) in contrast with measuring nerve spikes in the auditory system for representation of signal flow to the brain and internal representation of sound events useful for objective assessment of speeches and sounds.

Due to the complexity of the subsystems of the hearing system; the non-linearity and inherent feedback, spontaneous response and saturation at high stimulus levels, it is usually not possible to make use of analytical methods in modeling human auditory system. But the non-linearity and dynamic natures have been very useful for guiding the design of both hardware and computer simulation for the peripheral stage of auditory processingwhich is actualised with various electrical circuits and systems (James et al 2018; Karjalainen, 1987; Seneff, 1986).

Approaches at modeling the human auditory system were categorised into two namely: psycho-acoustical and physiological approaches (Kleczkowski, 1999). Psycho-acoustical approach normally does not involve complex functions at the higher stage of auditory system, while the physiological approach has very limited concerns for the physiology of the higher level of the human auditory system. Modeling human auditory system pertains majorly to the peripheral of the human auditory system and such models are categorised according to the segments of the hearing system. These include modeling of the external and middle ear, modeling the cochlea and modeling neural representations of speech signals.

Seneff (1986) noted that computationally modeling the auditory system starts from the acoustics of the external ear, particularly the acoustical details of the pinna and ear canals.

The following processes take place in humans perceiving and assessing the quality of speech or sound that they hear (Grancharov et al, 2006):

1. conversion of received speech signal into excitations that are conveyed by auditory nervesto the brain, and
2. cognitive processing of nervous excitations in the brain for necessary interpretations and actions.

#### 2.12.1.1 Flanagan's Auditory Model

The computational model for auditory mechanism proposed by Flanagan utilised physiological data obtained by Bekesy in his research work in the 1940's (Flanagan, 1962). The model shown in Figure 2.21 comprises of two parts based on the middle ear and the basilar membrane and Figure 2.22 being the block diagram representation.

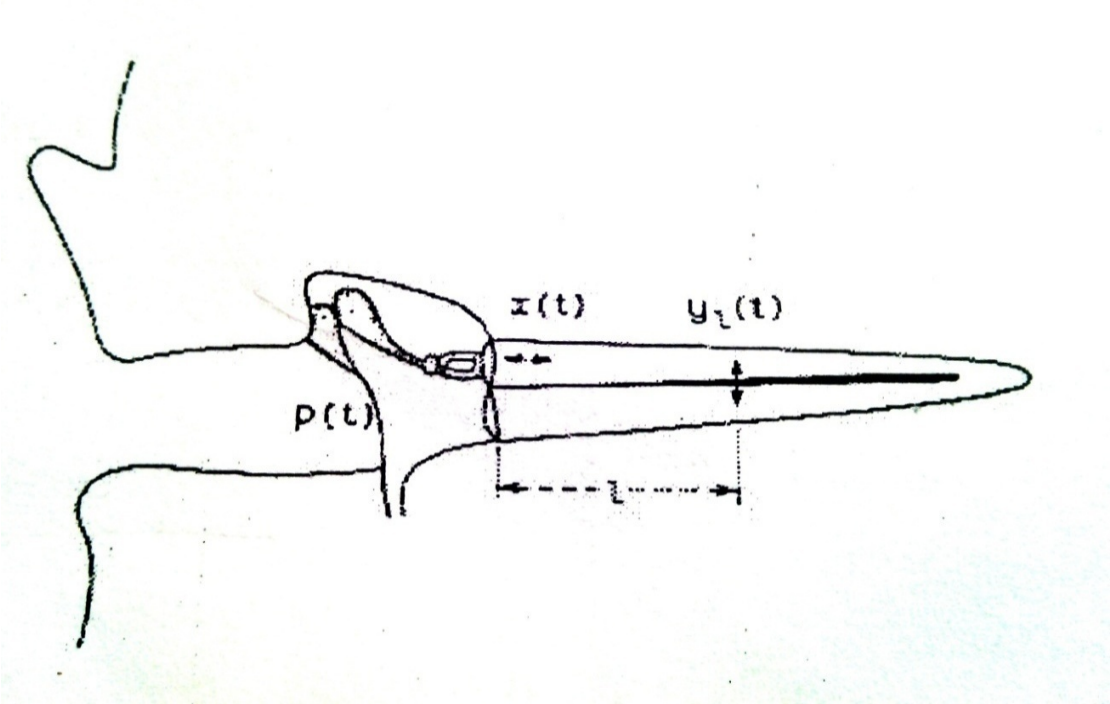



Figure 2.21. Components of the Flanagan Auditory Representation. (Source: Flanagan, 1962).

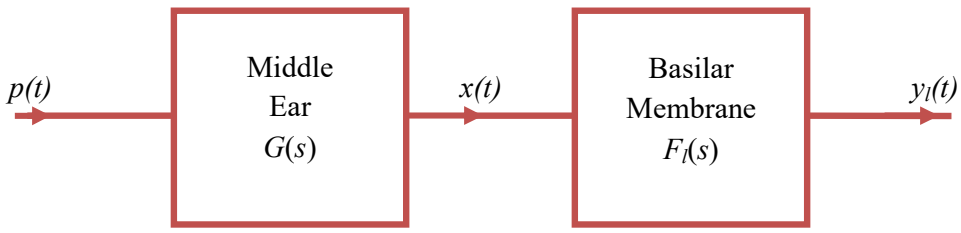



Figure 2.22. Block Representation of Flanagan's Model (Source: Flanagan, 1962).

The model was realized using lumped constant electrical circuits because of being modeled by rational functions, and has been a useful analytical tool for relating subjective behaviour of the auditory system and the acousto-mechanical operation of the human ear.

The parameter *p(t)* is the sound pressure impinging on the eardrum, *x(t)* is the eardrum output defined by displacement of the stapes, and $y_l(t)$ is displacement of the basilar at the distance $l$ from the reference point of the stapes. Analytical approximations to the relations between these quantities are given by Laplace functions, $G(s)$and $F_l(s)$ respectively. $G(s)$approximates middle ear transmission relating between $x(t)$ and $p(t)$, while $F_l(s)$approximates transmission from stapes to distance $l$ specified on the membrane. These relationships are given by the following equation:

$$\frac{Y_l(s)}{P(s)} = \frac{X(s)}{P(s)} \cdot \frac{Y_l(s)}{X(s)} = G(s) \cdot F_l(s) \qquad (2.18)$$

$G(s)$and $F_l(s)$ were fitted to existing physiological data, and given by:

$$G(s) = \frac{c_0}{(s+a)[(s+a)^2 + b^2]} \qquad (2.19)$$

where, $c_0$ is a positive real constant, $a$ and $b$ are the pole frequencies of the middle ear, $G(s)$, and are related by: $b = 2a = 2\pi(1500)$ rad/sec.

The inverse transform of $G(s)$defined as the stapes displacement response to the pressure of an impulse at the eardrum is given by:

$$F_l(s) = c_1 \left(\frac{2000\pi_{\ l}}{\beta_l + 2000\pi}\right)^r \beta_l{}^4 \left(\frac{s+\epsilon_l}{s+\gamma_l}\right) \left[\frac{1}{(s+\alpha_l)^2 + \beta_l{}^2}\right]^2 \times exp\left(\frac{-3\pi s}{4\beta_l}\right) \qquad (2.20)$$

where, $s = \sigma + j\omega$ is the complex frequency; $\beta_l$ is the radian frequency at the point of maximum response at distance$l$ from the stapes; $c_l$ is a constant value that specifies the proper absolute value of displacement; $\exp(-3\pi s/4\beta_l)$ is a delay factor of $3\pi/4\beta_l$ sec that brings the model's phase delay into line with the phase measured on the human ear; $[2000\pi\beta_l/(\beta_l + 2000\ \pi)]^r\beta_l{}^4$ is an amplitude factor which matches the

variations in peak response with resonant frequency $\beta_l$, as measure physiologically by Bekesy; the value of the exponent is taken as r = 0.8;$\gamma_l$, $\epsilon_l$, and $\alpha_l$are pole-zero constants appropriate to the point $l$ and are related to $\beta_l$.

Beyond the computational modeling done by Flanagan for studying the relation between the physiological and subjective auditory characteristics, simulation was also carried out with electronic circuits which have been applied in areas of pitch perception, binaural lateralization, threshold sensitivity and masking.

#### 2.12.1.2 Lyon's Model

Lyon's model is a multi-level algorithm for sound analysis carried out by modeling behaviour of the cochlea (the inner ear). It preserves important details in both time and frequency for robust sound analysis and it is suitable for processing of speech and other sounds on real-time basis (Lyon, 1982). Sub-models used in depicting how the basilar membrane and the organ of Corti behave are:

- linear and time-invariant filters,
- a nonlinearity detection given as half-waved, and
- a compression of wide dynamic range of mechanical domain into a range appropriate for neural representation with use of complex nonlinear mechanisms.

Based on the knowledge of the operation of the cochlea, Lyon developed an analog electronic cochlea using analog time-continuous CMOS VLSI technology incorporating variable-Q, near linear second-order filter cascades that simulates the fluid-dynamic travelling-wave system of the cochlea and the effects of adaptation and active gain of the outer hair cells (Lyon and Mead, 1988). Actions of low-passand band-pass filters were used in modelingsoundenergy propagation as hydro-dynamic waves in the fluid and partition system of the cochlea and the membrane velocity detected at each hair cell respectively. The filtering mainly actsto separate complex sound mixtures into regions of high SNR so that different frequencies are separated while as well preserving time resolution like separating responses into different pitch pulses (Lyon, 1982).

The Lyon's cochlea model made use of half-wave rectifier to model the movement of inner hair cells, since it only generates neural spikes when it is moved in one direction and does not when it is moved in the other direction. So, at the detection

stage, outputs of the filter stage are converted by amplitude demodulation using the diode amplitude modulator.

The activity of outer hair cells was modeled using cascaded Automatic Gain Controls (AGC). Experimenting on cochlea computational model, Lyon concentrated on the effects of nonlinear/time varying multichannel automatic gain control (MAGC) stages, with adaptation to a wide dynamic range of auditory stimuli (Lyon, 1986).
In summary, the Lyon's model block diagram and the signal flow between the component parts is shown in Figure 2.23.The data obtained from the cochlea model is known as cochleagram while correlogram is a summary of its periodic information, as shownby the graphical overview in figure 2.24. Figures 2.25 and 2.26 are the filter bank comprising of a cascade of 86 filters which feed the half-wave-rectifier (HWR), and the network of four AGC phases in cascade with the gain of the AGC'sdepending on the time constant from output samples of the adjacent channels and the overall output representing the neural firing rate (Hou et al, 2006). With a variety of detection thresholds and separate tuning curves built for the Multichannel Automatic Gain Control (MAGC), it was possible to play around with the outputs of the channels to demonstrate cochlea's responses to stimulus frequencies, as was viewed on an equal-loudness curve.

Lyon also investigated neural firing effects in computational modeling for auditory processing (Lyon, 1984). The work was focused more on research in speech recognition and hearing, and bothered on models of pitch perception, binaural directional perception, and sound separation.

#### 2.12.1.3 Meddis' Inner Hair Cell (IHC) Model

The Ray Meddis IHC model is a probabilistic model built around the IHC's physiology, which simulates activities of neurotransmitters released from hair cells in the cochlea and produce firing spikes for the neural fibres. The Meddis model is a very efficient computational model that is very appropriate for providing major input to bigger systems, for example, central-auditory processing and speech-recognition devices (Hewitt and Meddis, 1991).

Meddis highlighted that models which possess primary auditory fibre activity have existed in the past. Generated at their output are sequences of spikes in time domain, which in response to stimulating waveforms consist of two stages. These stages are: function that relates acoustic stimulation to transmitters that are released

from hair cells into the synaptic cleft and the characterising of auditory nerve fibre response to the presence of the transmitter released into the cleft (Meddis, 1986; Zilany et al, 2014).

As shown in Figures 2.27 and 2.28 (McEwan and Schaik, 2002), in the Organ of Corti on top of the basilar membrane are many micro organs. Major among these organs are the inner hair cells, which act as transducers that convert the mechanical

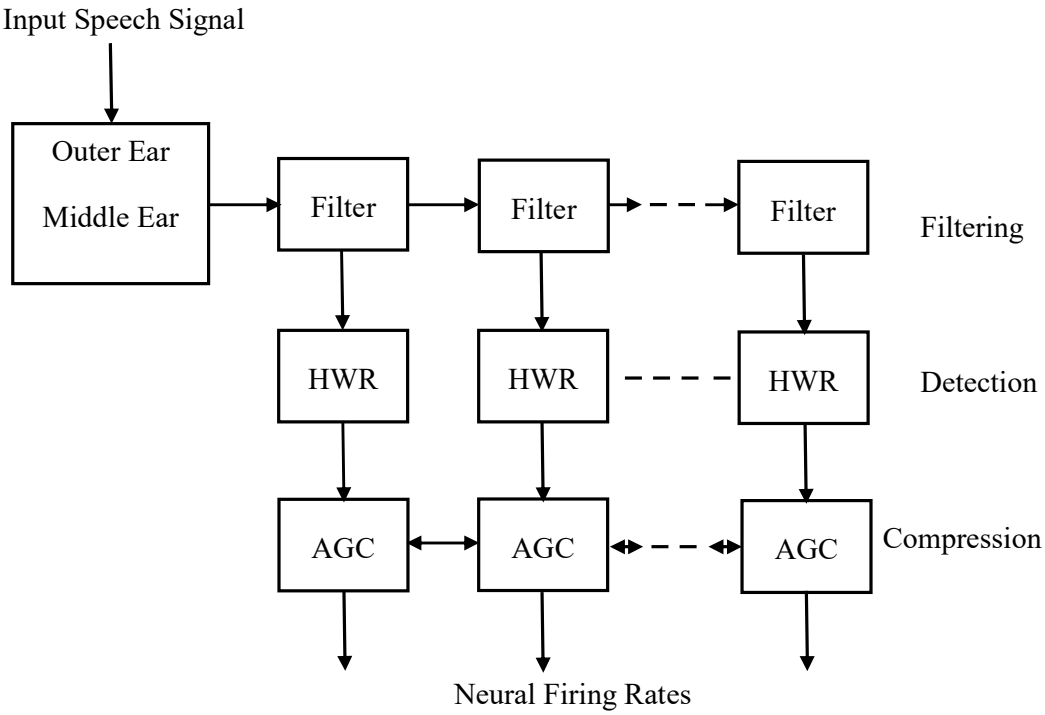



Figure 2.23. Component Flow Diagram of the Lyon's Model. (Source: Lyon, 1986).

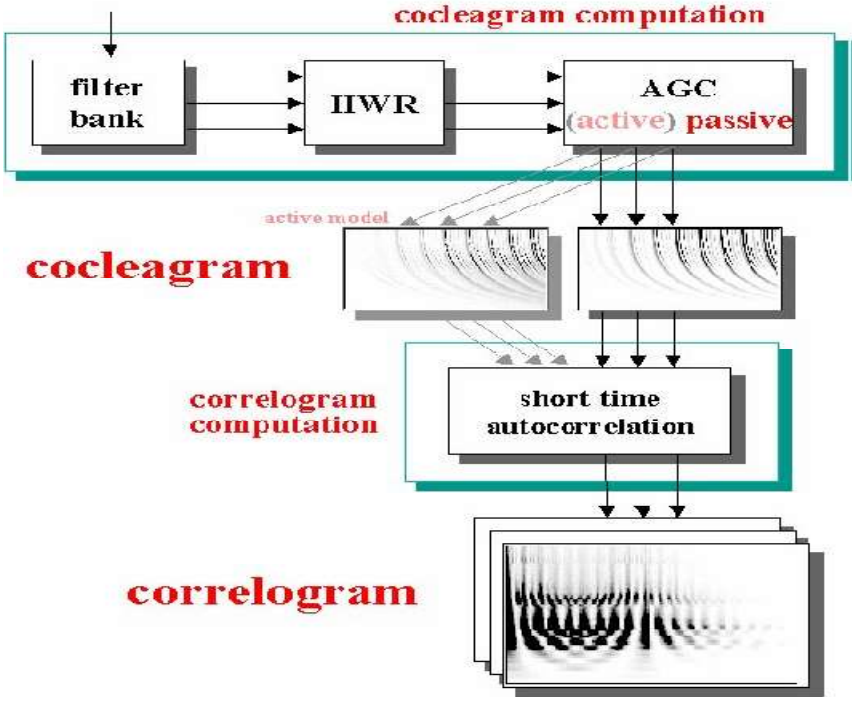



Figure 2.24. Graphical Overview of the Lyon's Auditory Model. (Source: Hou et al, 2006).

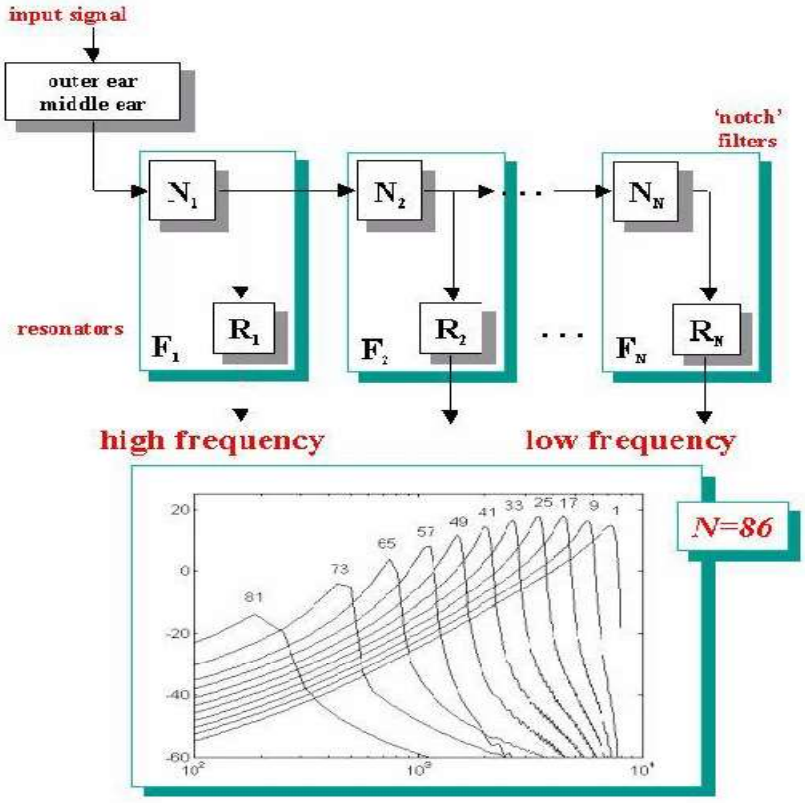



Figure 2.25. Cascaded Filter Bank of the Lyon's Auditory Model. (Hou et al, 2006).

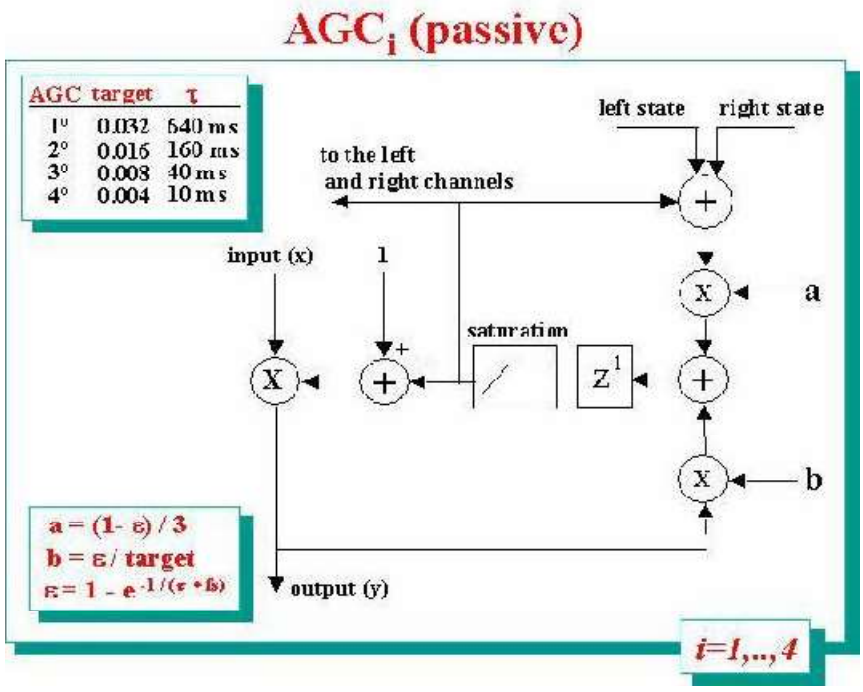



Figure 2.26. Four AGC Phases Cascaded to the Output of the Model. (Hou et al, 2006).

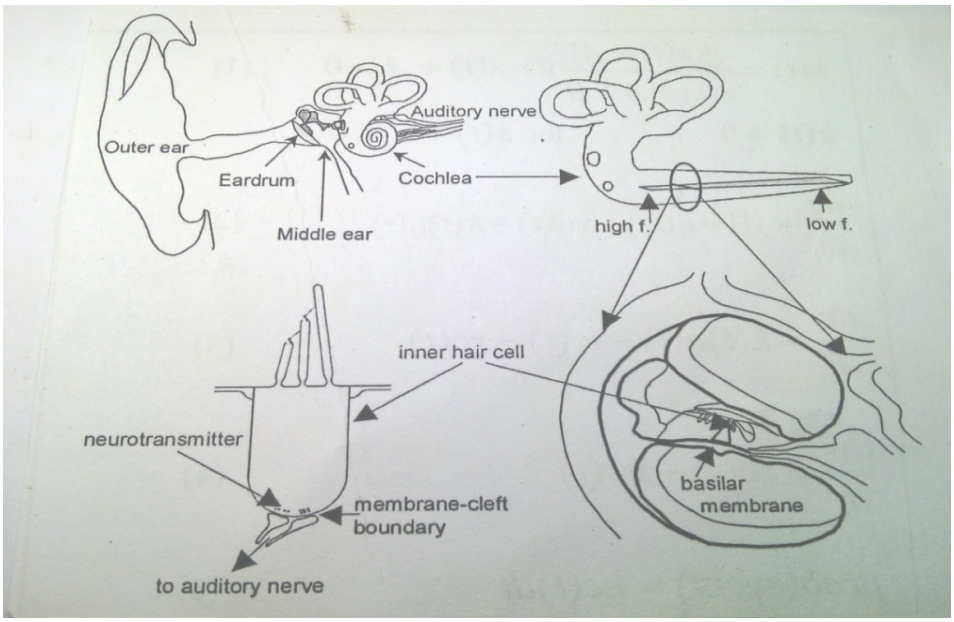



Figure 2.27. Human Ear showingInner Hair Cell (Source: McEvan and Schaik, 2002)

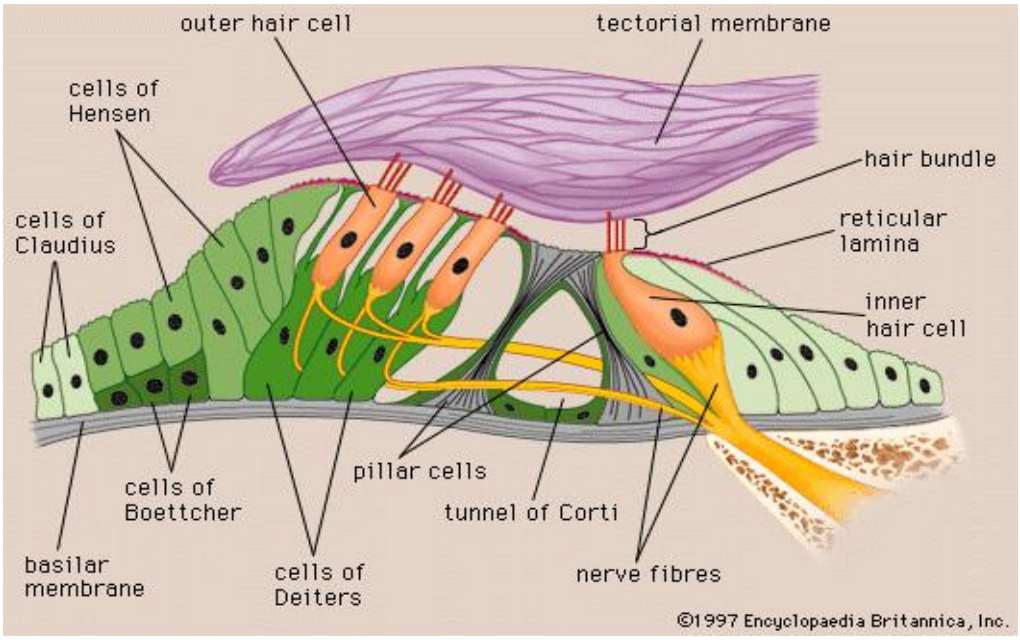



Figure 2.28.IHC in the Organ of Corti (Source: McEwan and Schaik, 2002)

stimulus of sound to firing spikes for the auditory nerve cells (Bruce et al, 2018).

The neurotransmitters are transferred between three reservoirs shown in Figure 2.29, namely: the factory that is releasing neurotransmitters at the boundary of the hair cell and delivering them to the pool of free transmitters, which releases them into the cleft. The third reservoir reprocesses and stores transmitters that never get to the cleft and are diverted back to the cell. Quantity of neurotransmitters leaving the pool into the cleft varies based on changes in permeability of cell membrane as a function of intra-cellular voltage and magnitude of mechanical stimulus. This means that some transmitters never get beyond the cleft and are diffused into the cell.

Transmitters in the cleft stimulate the afferent fiber of an auditory nerve cell to produce firing spikes in it. The probability of nerve firing spikes produced is determined by the amount of transmitter that is in the cleft (Meddis and Lopez-Poveda, 2010).

In examining the statistics of neural firing spikes Summer et al (2002) highlighted studies that have been done on the relationship between the mean firing rate and its variance, noting that at higher firing rates the variance is less than the mean indicating a departure from the predictions of a Poisson model of spiking statistics.

Derived from the processes of the neurotransmitters are the equations representing the Meddis IHC model, the following were stated (Zhang, 2005; Meddis, 1986; McEwan and Schaik, 2002):

$$k(t) = \begin{cases} \frac{g[s(t)+A]}{s(t)+A+B} \; for\; s(t) + A > 0 \\ \\ 0 for\; s(t) + A < 0 \end{cases} \quad (2.21)$$

$k(t) = gA/(A + B)$ – in the absence of sound (2.22)

$$\frac{dq}{dt} = y[1 - q(t)] + xw(t) - k(t)q(t) \quad (2.23)$$

$$\frac{dc}{dt} = k(t)q(t) - lc(t) - rc(t) \quad (2.24)$$

$$\frac{dw}{dt} = rc(t) - xw(t) \quad (2.25)$$

$$prob(event) = hc(t)dt \quad (2.26)$$

where, A, B, and g are model constants, $k(t)$ is permeability of the cell membrane, $k(t) = gA/(A + B)$ is spontaneous hair cells' response at rest, $q(t)$ is the measure of neurotransmitters present in the pool, $y[1 - q(t)]$ is the amount of neurotransmitters

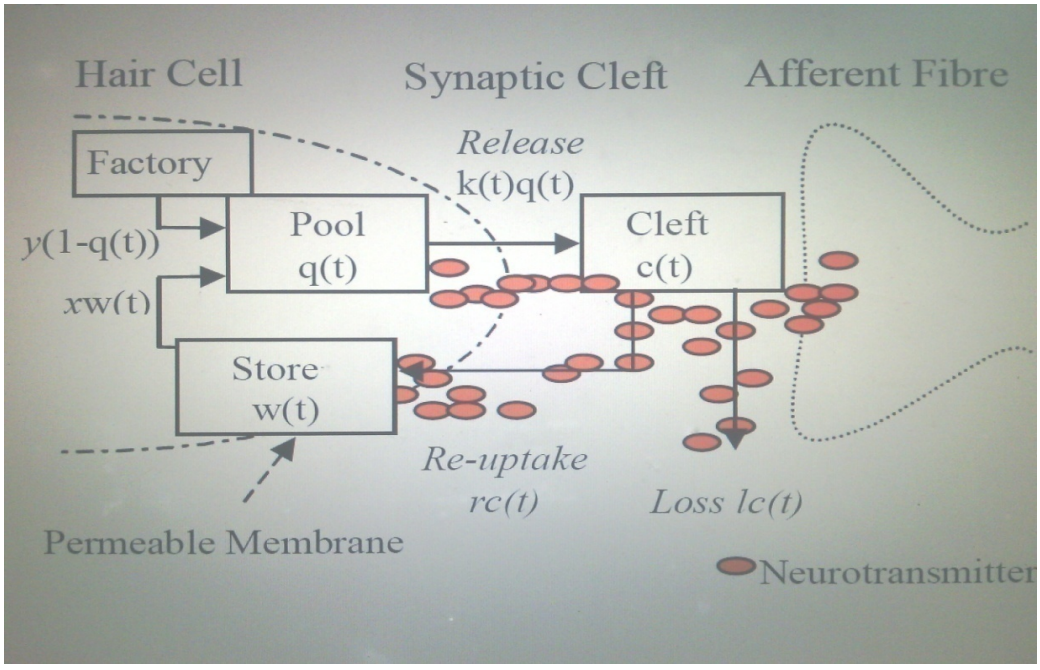



Figure 2.29. The Meddis IHC Model.(Source: Meddis, 1986).

produced, $xw(t)$ is the quantity of neurotransmitters reprocessed, $k(t)q(t)$ is the quantity of neurotransmitters in the cleft, $lc(t)$ quantity of neurotransmitters lost through diffusion, $rc(t)$ is the quantity of neurotransmitters actively fed back to the reprocessing store, $r$ is the rate at which reprocessing store receives neurotransmitters and $x$ the rate at which it returns it to the free transmitter pool.

For simulation and implementation, the Meddis model shows in Figure 2.30 the transformation to electrical current domain. This made it possible to make use of log domain filters in designing and building VLSI circuits to realize the model, whereby $I_{max}$ is a constant current entering the multiplier (MULT), *ga*, *gb*, *gc*, *gd* are variable gain current mirrors, $I_q$ is the feedback current to the multiplier, $I_{stim}$ is the stimulus current from the inner hair cell, and $I_{max}$ constant current entering the multiplier (Freedman et al, 2014; McEwan and Schaik, 2002; McEwan and Schaik, 2003).

In Figure 2.30, the Meddis circuit consists of a Half Wave Rectification (HWR) function, a Multiplier (MULT), four variable gain current mirrors (*ga*, *gb*, *gc*, *gd*), and three first order low-pass filters, (CLEFT, STORE and POOL). First order trans-linear or log-domain low-pass filters implement the time constants, while multiplication of the stimulus and feedback ($I_q$ , $I_{stim}$) is carried out by a trans-linear multiplier, and the output by the constant current $I_{max}$ is normalized.

Current domain equivalent equations of the Meddis equations stated earlier are given by:

$$hwr(V_{in}) \approx I_{stim} = \frac{I_{bias}}{1+ e^{(V_{ref}-V_{in})/nU_T}} - I_{shift} \tag{2.27}$$

$$I_d = I_q k(t) = \frac{g \times I_{stim}}{I_{max}} I_q \tag{2.28}$$

The Meddis IHC model was adopted in building a model for objective prediction of speech quality (Hauenstein, 1998). In doing so, he compared the specific-loudness pattern internal signal representation of the reference signal, $x(k)$, and of the distorted (output) speech signal, $y(k)$ with the generation of auditory nerve spikes as an alternative internal signal representation. During signals preprocessing, codec's gain and delay were measured and compensated. A VAD similar to that used in GSM networks was adopted for eliminating redundant speech pauses, and FIR-bandpass filters were used to model the output and to approximate the frequency response of the telephone device.

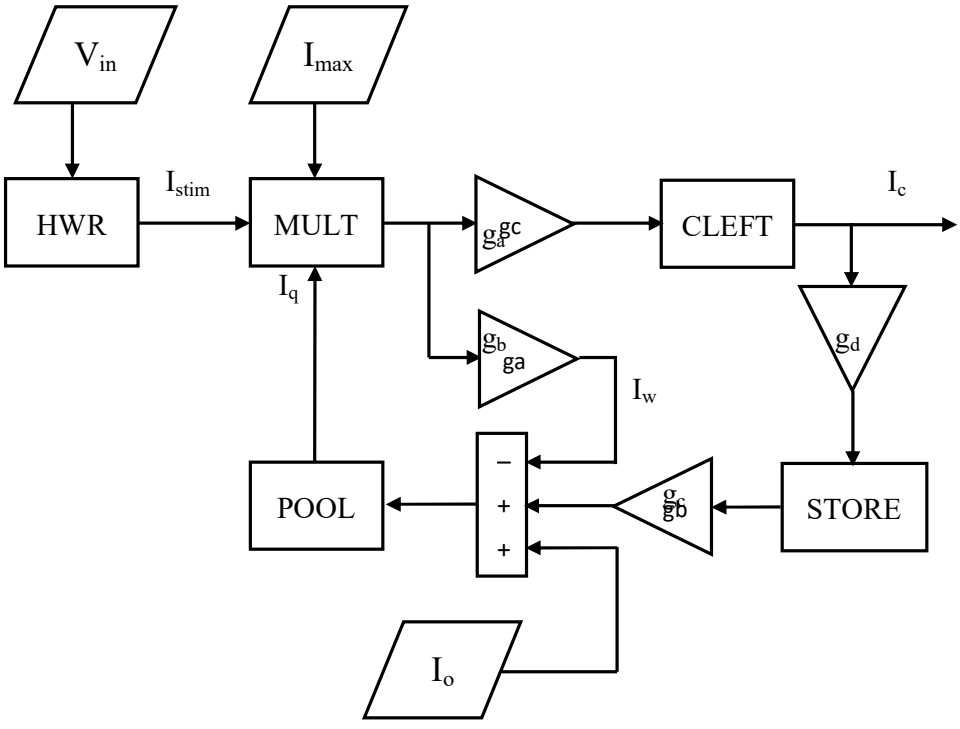



Figure 2.30. The Electrical Current Mode of the Meddis model (McEwan and Schaik, 2003)

The firing pattern of the preprocessed signals ($x'$ and $y'$) were then calculated by adopting the Meddis IHC model to obtain the internal representations ($p_{xv}(\kappa)$ and $p_{yv}(\kappa)$) of the signals from which a distance measure ($d(\kappa)$) for calculating the quality of the transmitted speech was obtained from the following equations:

$$d(\kappa) = \alpha\, d_{||}^{+}(\kappa) + (1 - \alpha) d_{||}^{-}(\kappa) \quad (2.29)$$

given $$d_{||}^{+}(\kappa) = \sum_v max\{[p_{yv}(\kappa) - p_{xv}(\kappa)], 0\} \quad (2.30)$$

and $$d_{||}^{-}(\kappa) = \sum_v max\{[p_{xv}(\kappa) - p_{yv}(\kappa)], 0\} \quad (2.31)$$

where $d_{||}^{+}(\kappa)$ is the mean absolute error of the firing probabilities of $y''$whichis higher than those of $x''$, while $d_{||}^{-}(\kappa)$ is that of those which are less than of $x''$. Therefore, $d(\kappa)$ is the weighted sum of $d_{||}^{+}(\kappa)$ and $d_{||}^{-}(\kappa)$.

A mean distance $\bar{d}$ was obtained from averaging the values of $d(\kappa)$ and monotonically mapped to a MOS estimator. Three of the four tests conducted have correlations of the subjective and objective scores as 0.91, 0.94 and 0.95 respectively.

The human Inner Hair Cell/Auditory Nerve (IHC/AN) was modeled in real-time applications based on the use of SpiN Naker machine and a data transmission parallelism architecture wherein it was estimated about 30,000 AN fibres feed the auditory brainstem from each cochlea (James et al, 2018). This was used to simulate a real-time full-scale digital model of the human auditory pathway.

#### 2.12.1.4 The Auditory Image Model (AIM)

AIM was originally developed by Roy Patterson and his team in 1991, for the purpose of analysing common sounds like music and speech.The model builds on the versatile and popular Meddis modeland it is extensively described in Patterson and Holdsworth (1991) and Patterson et al (1995).

As a software-based modular architecture,itcontains two modules for simulating auditory spectral analysis, neural encoding and temporal integration, namely: functional and physiological modules, shown in Figure 2.31. It also contains new forms of generating auditory images that can be replayed to produce multiple auditory perceptions useful toexplainauditory system'sdynamic response to common sounds.

The bold types in Figure 2.31 are the functions for each module; the rectangular boxes are the implementation, while the italic types are the simulation.Users are allowed to move from functional to physiological version and vice versa.

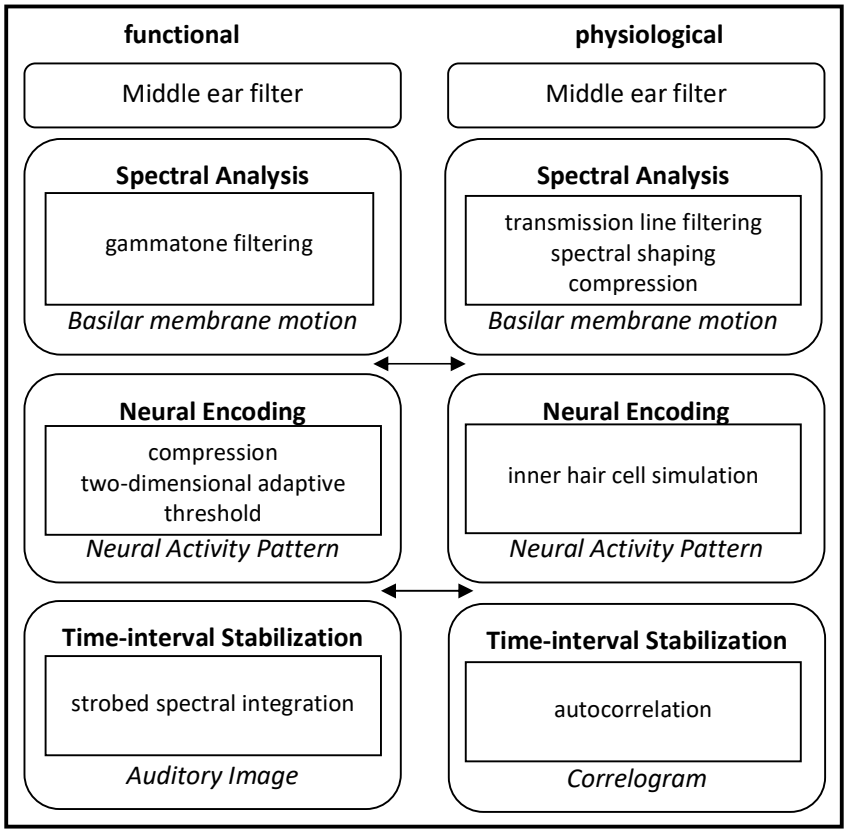


Figure 2.31. Three-stage Structure of the AIM Modular Architecture. (Source: Patterson et al, 1995).

The spectral analysis of sound waves is carried out using banks of auditory filters to convert sound waves into model that represents motion of the basilar membrane. Under the functional framework, the auditory filters used are linear, gammatone filters, while under the physiological framework nonlinear transmission line filters are used. The gammatone auditory filters are defined bythe impulse response given in time-domain by (Patterson et al, 1992):

$$gt(t) = \text{a}\ t^{(\text{n-1})} \exp(-2\text{p}bt) \cos(2\text{p}\ f_c t + \emptyset) \qquad (t > 0) \tag{2.32}$$

where, the primary parameters of the filter, b is the impulse response duration and the filter bandwidth, $n$ is the order of the filter and $f_c$ is the center frequency.

The human data of the auditory filter is summarized with the Equivalent Rectangular Bandwidth (ERB) given by:

$$ERB = 24.7[(4.37f_c/1000) + 1] \tag{2.33}$$

This function is referred to as the same with the 'cochlea frequency position' function and is the physiological basis for the 'critical band' function with the 3-dB bandwidth of the gammatone filter being 0.887 multiplied by the ERB. Equation 2.33 indicates that the auditory frequency resolution as described by ERB is approximately constant-Q at high frequencies (>2 kHz) and that using ERB units, the range of audible frequencies is discretized as a bank of 39 adjacent filters whose ERB number is given byNecciari et al (2013):

$$ERB_{\text{num}}(f_c) = 9.265\ \text{In}\left(1 + \frac{f_c}{228.8455}\right) \tag{2.34}$$

where $f_c$is the center frequency. Equation 2.34 is the required ERB scale for plotting psycho-acoustical data on a perceptual frequency axis. The neural encoding module simulates mechanical/neural transduction process similar to what happens in the Meddis IHC model. It also stabilizes the Basilar Membrane Motion (BMM) and

converts it into a Neural Activity Pattern (NAP) in the auditory nerve. In doing this, the motion of the basilar membrane is converted by the IHC's into neural transmitter. This stage also has the two alternative means for generating the NAP, namely: the bank of adaptive threshold units, that rectifies and compresses the BMM and applied adaptation and suppression in time and frequency respectively; and the bank of inner hair cell simulators, similar to the Meddis IHC model simulation.

AIM is coded with software packages that convert sound waves into auditory images, with computational versions in MATLAB and C programming languages known as AIM-MAT and AIM-C respectively. The AIM-MAT has a Graphical User Interface (GUI) that enables one to investigate auditory processing stage by stage, while AIM-C is a real time version that is suitable for batch processing of sound databases. The C code is obtainable for installation and compilation as a compressedarchive from ftp.mrc-apc.cam.ac.uk(Patterson et al, 1995). AIM in MATLAB was extensively described and analysed by Stefan et al (2004). MATLAB was used to code the processing modules, resource files and the GUI with details of how to use it contained in http://www.mrc-cbu.cam.ac.uk/cnbh/aimmanual.

### 2.12.2 Auditory-Based Intrusive Speech Quality Assessment Models

Intrusive speech assessment models are built upon mimicking the human auditory system. Speech signals are transformed into auditory nerve excitations through psychoacoustic processes ofBark scale frequency warping and conversion of spectral power to subjective loudness through the use of different psychoacoustic models (Grancharov and Kleijn, 2008). This is followed by cognitive processing of extraction of compact features from the auditory excitations and combining them to give a picture of the perceptual speech quality.

Beyond the earlier waveform–comparison and spectral related algorithms, are the more recent perceptual-domain related model algorithms, some of which met the standardisation efforts by ITU-T. These include Bark Spectral Distortion (BSD), Modified Bark Spectral Distortion (MBSD), Perceptual Speech Quality Measure (PSQM) standardised in ITU-T Rec. P.861(1996), Perceptual Analysis Measurement System (PAMS), Perceptual Evaluation of Speech Quality (PESQ) standardised as ITU-T P.862(2001), and Perceptual Objective Listening Quality Analysis (POLQA).

#### 2.12.2.1 Perceptual Evaluation of Speech Quality (PESQ)

PESQ is an intrusive objective technique of automated E2E assessment of speech quality of speech codecs and telecommunication networks (Conway, 2004).The narrow-band version was standardised as ITU-T Rec. P.862 in 2001 with the wideband version standardised as ITU-T Rec. P.862.2 in 2007. The wideband version covered wideband audio systemsof frequency from 50-7000 Hz, andcould also be adopted for systems with a narrower bandwidth. It allows for increased speech quality and intelligibility.

PESQ modelis a "full reference" algorithm in that it requires the reference speech signal in estimating the quality of the degraded speech signal. It was designed ultimately for speech enhancement and transmission systems, while for audio enhancement and transmission systems, the Perceptual Evaluation of Audio Quality (PEAQ) algorithm was designed and standardised asITU-T Rec. BS.1387-1 (2001). The PEAQ model also makes it possible to evaluate the quality of stereo signals (Schafer et al, 2013).

PESQ is extensively adopted in carrying out series of voice quality assessments on all networks. It is also being used on Voice-over IP (VoIP) networks and for predicting speech quality in modern codecs (Sun and Ifeachor, 2006). It offers high accuracy and repeatability particularly in dedicated tests of speech quality in live telecommunication networks, like in drive tests on mobile networks.

The PESQ algorithm is widely used by manufacturers, vendors and operators of telecommunication equipment and networks for speech quality testing. The PESQ algorithm has been deployed in applications like in the development of new coding algorithms for speech signals, for exploring quality effects of variations of bit rate, input levels, and channel errors of speech codecs. It has also been deployed in equipment selection for comparing the quality effects of distortion scenarios on communication systems and technologies, and in equipment and network monitoring and optimization (Psytechnics, 2004).

Shown in Figure 2.32is the structure of PESQ model (Kondo, 2012; Rix et al, 2001), in which the original or reference and degraded speech signals are level-aligned to the same power level, filtered (FFT), time-aligned, equalized, and processed through auditory transformation. In time-aligning the original and degraded speech signals, delays existing between their segmented parts are calculated and the algorithm make efforts at aligning the two signals. Using a perceptual model built around perceptual frequency (Bark) and loudness, internal representation of the

signals are formed which enable the psychophysical distance between them to be obtained for estimating speech quality degradation (Miroslav and Rozhon, 2012).

Input speech signals are broken into phonemes each of 32ms duration from which spectral characteristics are calculated and perceptual differences from the reference signal are obtained for each phoneme (Kajackas and Vindasius, 2010). These are the distortion parameters which are extracted, aggregated and mapped to the subjective MOS.

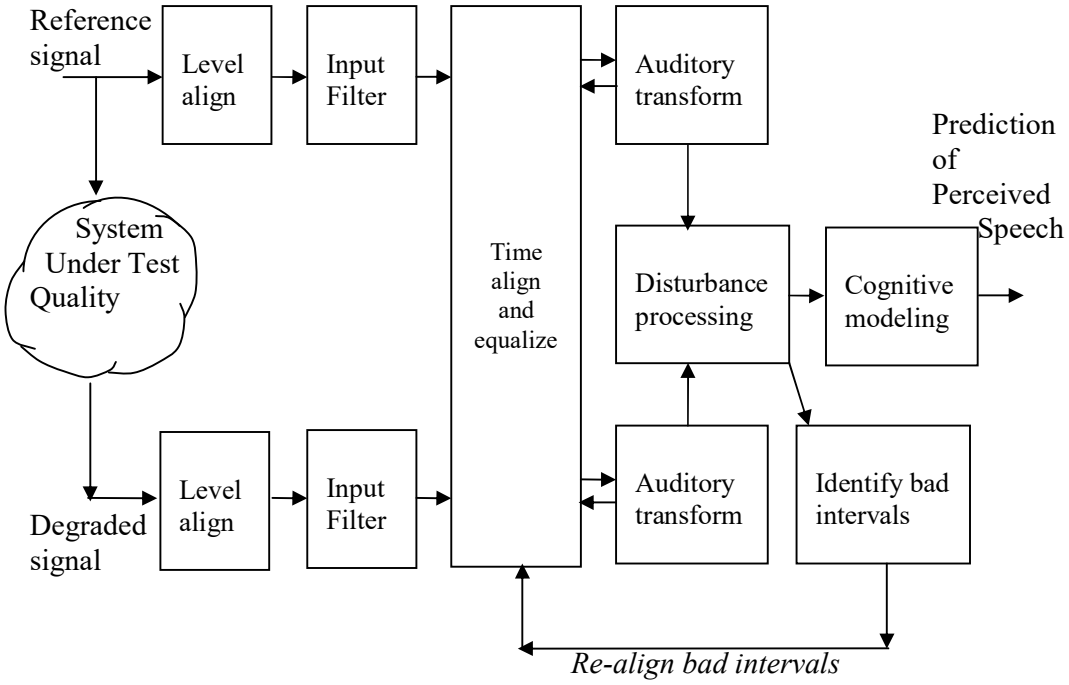



Figure 2.32. Structure of the PESQ Model. (Source: Rix et al, 2001).

The auditory transform in PESQ algorithm maps the processed signals with loudness perception as a representation of the human auditory system. It includes the use of Fast Fourier Transform (FFT) with Hamming window to calculate the instantaneous power spectrum (Bark spectrum) in each frame. The mean Bark spectrum of the active speech frames was calculated to obtain a ratio between reference and degraded spectra and used for equalising the reference to the degraded at $\pm$ 20dB. The gain variation of the reference and degraded spectra is determined, and mapping of the Bark spectrum to loudness is done to obtain the perceived loudness in time–frequency representations.

The values of differences in disturbance inherent in the reference and degraded speeches are aggregated to obtain a non-linear average given by Rix et al(2001):

$$L_p = \left[\frac{1}{N}\sum_{m=1}^{N} disturbance(m)^p\right]^{1/p} \quad (2.35)$$

To calculate the difference existing between the reference and degraded speech signals, these speeches are broken into phonemes of about 32ms whichare overlapped by 50% and 20 phonemes are aggregated into a long 320ms syllable.

A measure of perceived disturbance is estimated for each phoneme by calculating the symmetric and asymmetric disturbances $D_{sn}$ and $D_{an}$ and the aggregation of phoneme disturbances $D_{sn}$ and $D_{an}$ for every syllable, $L_{DS}$ and $L_{DC}$ are given by Kajackas and Anskaitis (2009):

$$L_{DS} = \left(\frac{1}{20}\sum_{i=1}^{20} D_{sn}^6\right)^{1/6} \quad (2.36)$$

and,

$$L_{DC} = \left(\frac{1}{20}\sum_{i=1}^{20} D_{an}^6\right)^{1/6} \quad (2.37)$$

The aggregated symmetric and asymmetric syllable disturbances are obtained by mean square algorithm and given by:

$$d_{sym} = \left(\frac{1}{N}\sum_{i=1}^{N} L_{DS}^{2}(i)\right)^{1/2} \quad (2.38)$$

and,

$$d_{asym} = \left(\frac{1}{N}\sum_{i=1}^{N} L_{DC}^{2}(i)\right)^{1/2} \quad (2.39)$$

where, *N* is the number of syllables in PESQ measurement window *T*.

The speech quality prediction made from the two disturbance parameters, given above, is given by Rix et al (2001), and Kajackas and Anskaitis(2009):

$$PESQ_{MOS} = 4.5 - 0.1\, d_{SYM} - 0.0309\, d_{ASYM} \quad (2.40)$$

It was discovered that PESQ had better correlation to subjective scores than all previous auditory-based (intrusive) speech quality prediction algorithms: on the average, 0.962 to 0.924 in PSQM and 0.883 in MNB for the same speech samples.

#### 2.12.2.2 Perceptual Objective Listening Quality Assessment (POLQA)

While PESQ model remains the state-of-the-art voice quality assessment technique, POLQAas a new algorithm was developed and standardized as ITU-T Rec. P.863, as anext-generation voice quality assessment technique. It was developed for predicting speech quality from narrow to wideband and super-wideband (50 – 14,000 Hz), for high density (HD) voice, and for evaluation, optimization and monitor of the voice quality on next-generation networks (ITU-T Rec. P.863, 2014; Sloan et al, 2017) . It is still notwithstanding undergoing extensive testing and reviews.

### 2.13 Review of Previous Works on Intrusive Objective Assessment of Speech Quality using PESQ Algorithm

Although PESQ is a robust algorithm and has been in use for a couple of years for estimating the quality of transmitted and processed speech signals and for the optimization of telecommunication networks, a number of research efforts have been on-going to address limitations and constraints that require improvements in the original PESQ algorithm.A number of critical evaluations, improvements, modifications and reviews have beingachievedby researchers. These include efforts at correcting time and level alignment problems, signal spectrum mismatch, mapping

from the arbitrary PESQ score to the PESQ MOS-LQO score before correlation with subjective MOS. Also, modification of PESQ was achievedat improving performance of estimation of speech quality in low rate codec (less than 4 kbits/s) (Rix et al, 2006).

### *1. The Works of Zhang et al, 2013; and Zhang et al, 2014*

Noting the fact that the ITU-T PESQ algorithm (ITU-T Rec. P.862 & P.862.2) made use of the Bark-scale frequency, (Zhang et al, 2013) in their work decided toreplace the Bark scale with the ERB scale and the Moore and Glasberg loudness model as against Zwicker loudness model used in ITU-T PESQ algorithm. They claimed that ERB scale is more accurate than Bark scale for the description of the frequency selectivity of the human auditory system at lower frequencies.

The ERB scale is built upon the differential equation of the center frequency, *f*, of the human auditory filter, given by:

$$\frac{df}{dv} = 6.23f^2 + 93.39f + 28.52 \tag{2.41}$$

Solution of this differential equation produces an expression of frequency, in Hz, given by:

$$f = \frac{676170.4}{47.06538 - e^{0.08950404v}} - 14678.49 \tag{2.42}$$

where the ERB value, $v$, obtained as a subject of expression, is given by:

$$v = 11.17268\, log\left(1 + \frac{46.06538f}{f + 14678.49}\right) \tag{2.43}$$

With these claims and replacements, they came up with an improved version of PESQ algorithm which they called the New Perceptual Evaluation Speech Quality (NPESQ). They validated their work on three different wireless technology codecs, namely: Adaptive Multi Rate (AMR) codec, Enhanced Variable Rate Codec (EVRC) and Enhanced Variable Rate Codec – B (EVRC-B). For closeness of fit between the subjective tests and the quality score of the original PESQ model, they obtained a correlation coefficient of 0.8565 using AMR codec and 0.9335 with their NPESQ. Using EVRC codec, they obtained a correlation coefficient of 0.8985 with normal PESQ and 0.9390 with their NPESQ, while on EVRC-B, they obtained correlation coefficient of 0.8978 with normal PESQ and 0.9125 with their NPESQ.

With their newly formulated NPESQ having better coefficient correlation to subjective MOS scores than the normal ITU-T PESQ, they proved that their NPESQ

is more accurate and novel as an objective speech quality assessment algorithm than the ITU-T PESQ algorithm

### *2. The Work of Shiran and Shallom, 2009*

In evaluating performance of the PESQ algorithm at the time alignment stage,it was noted that it could not align continuous variable delays in the speech signals particularly signals that the rate of packet loss is high and for which dynamic time processing is exhibited. This is because of its piecewise constant delay estimation.

The authors developed a new algorithm to align the time and by so identify both fix and variable delays in the reference and degraded speech signals through the use of Dynamic Time Warping (DTW) in place of utterance correlation and splitting methods used in the original PESQ algorithm. They adopted dynamic programming in evaluating similarities between the reference and degraded speech signals. This involves time-registering of the signals so as to correctly match time-aligned pattern vectors calculated for the signals. Then, an optimal path, *P*, was found for the pairs of reference and test pattern vectors, $m(k)$ and $n(k)$ given by:

$$P = \{m(k), n(k)\}_{k=1}^{K} \quad (2.44)$$

where $k = 1 \ldots K$ is the common time scale, and a minimal distance function calculated to optimize the path.

The result of what they tagged Enhanced PESQ showed some improvements over the original PESQ algorithm when Deutsche-Telecom (DT) speech database was used. The Pearson's correlation coefficient after mapping for EPESQ was 0.901 against 0.881 for PESQ and Root-Mean-Square Error (RMSE) of 0.219 against 0.228 for PESQ for narrowband. For wideband signals, the correlation coefficient for EPSEQ was 0.822 against 0.818 for PESQ and the RMSE was 0.366 against 0.372 for PESQ.

### *3. The Work of Hu and Loizou, 2008*

It was argued that many of objective techniques developed over time for estimating speech quality by assessing distortions suffered during coding and transmission of speech signalsmay not have been good enough to estimate the quality of speech enhancement carried out by noise suppression algorithms (Hu and Loizou, 2008). These researchers made use of 1792 processed speech samples of NOIZEUS

speech database and a couple of objective measures among which wasPESQ to estimate the quality of noisy speech enhanced by noise suppression algorithm from the perspectives of signal and noise distortions, and overall quality.

They noted that computing PESQ score is from a linear combination of symmetrical and asymmetrical disturbances, $D_{ind}$ and$A_{ind}$, weighted by coefficients in the following equation:

$$\text{PESQ} = a_0 + a_1 D_{ind} + a_2 A_{ind} \quad (2.45)$$

where, the parameters originally given as $a_0 = 4.5$, $a_1 = -0.1$ and $a_2 = -0.0309$ were optimized for quality of transmitted speech over networks.

In working with PESQ and in attempt at optimizing it, they came up with a modified version of PESQ. In it the above parameters were optimized for each of the quality conditions of signal and noise distortions, and overall quality by multiple linear regression analysis method making PESQ score to correlate well with them.

In correlating the quality scores obtained for objective measures to the subjective quality score, the estimated correlation coefficient obtained for PESQ for overall quality was ρ = 0.65, for noise distortion was ρ = 0.57 and for background distortion was ρ = 0.48. With the modified PESQ the correlation coefficient improved to an average of ρ = 0.89 for overall quality, ρ = 0.81 for noise distortion and ρ = 0.76 for background distortion.

### *4. The Work of Malfait et al, 2008*

In the preprocessing of original and degraded speech signals by the PESQ algorithm, they are first taken through time alignment frame-by-frame. Majoring on this first stage in the PESQ algorithm, (Malfait et al, 2008) noted that a large error in the quality score could result from a few misaligned frames resulting in poor correlation with the subjective scores. They therefore experimented with reengineering the PESQ time alignment in order to attain a near perfect delay profile.

First, they pre-aligned the speech signals; manually adjusted them to ensure delays in them are accurately stated and compared the PESQ result with the subjective score each time progressively capping the maximum PESQ alignment error until a near perfect alignment was obtained. For a misalignment of 10ms, they obtained a correlation of 0.93 with the subjective score, a correlation of 0.973 for misalignment less than 5ms and have no significant improvement in the correlation coefficient for

misalignment down to about 1ms. They concluded that a time alignment of ±5ms seemed good enough for correct assessment of time-warped signals.

## 2.14 Review of Mapping Functions for Quality Estimation

Despite being a robust quality estimation technique, considerable and consistent efforts have been on-going on ways and measures atimproving various aspects of the objective speech quality estimation algorithm due to defects and constraints noticeable in it. These are to allow forbetter and more accurate estimateof the quality scores of transmitted speech signals. One of the major areas of constraintsin the PESQ model has been in the mapping of raw PESQ score to the generic subjective MOS. Figure 2.34 shows the mapping of raw PESQ score to yield the MOS forobjective listening quality (MOS–LQO) estimation.

This area of mapping the raw PESQ score to an accurate quality score rating has so far received very little research attentions as noted in its sparse reportage in literature. Accurately mapping of raw quality scores is a very important aspect of speech quality estimation in order to provide true indication of the quality of degraded speech. The raw PESQ output score which range between -0.5 to 4.5 is out of scale with the subjective quality scale of 1 to 5 which is the standard and generic quality scores for all objective speech quality measures.

This scale mismatch makes the objective speech quality inapplicable for direct qualitymeasurement of degraded speeches, and necessitated transforming from the raw objective (PESQ) score to its MOS-related score (MOS-LQO). Some of the mapping functions hereby reviewed has different levels of inaccuracies. There is therefore the need to develop more accurate mapping functions and this forms a major challenge of this research efforts.

### *1. Mapping Function for MNB*

An earlier objective speech quality measure known as Measure Normalization Blocks (MNB) and developed by Voran (1998) had the logistic function given by:

$$\mathrm{L(z)} = \frac{1}{1 + e^{az+b}} \qquad (2.46)$$

where z is the perceptual distance values of the MNB measure. It was noted that when $a > 0$, L(z) is a decreasing function of z. This is simply primitively symmetrical.

### *2. The PESQ NB ITU-T Rec. P.862.1 mapping function*

The ITU-T Rec. P.862.1 mapping function was developed as addendum to the ITU-T Rec. P.862 PESQ standard and given by:

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{(-1.4945 \quad 4.6607)}} \tag{2.47}$$

This mapping function was trained on both simulated and field-collected speech samples and has been widely used for estimating the MOS for listening quality

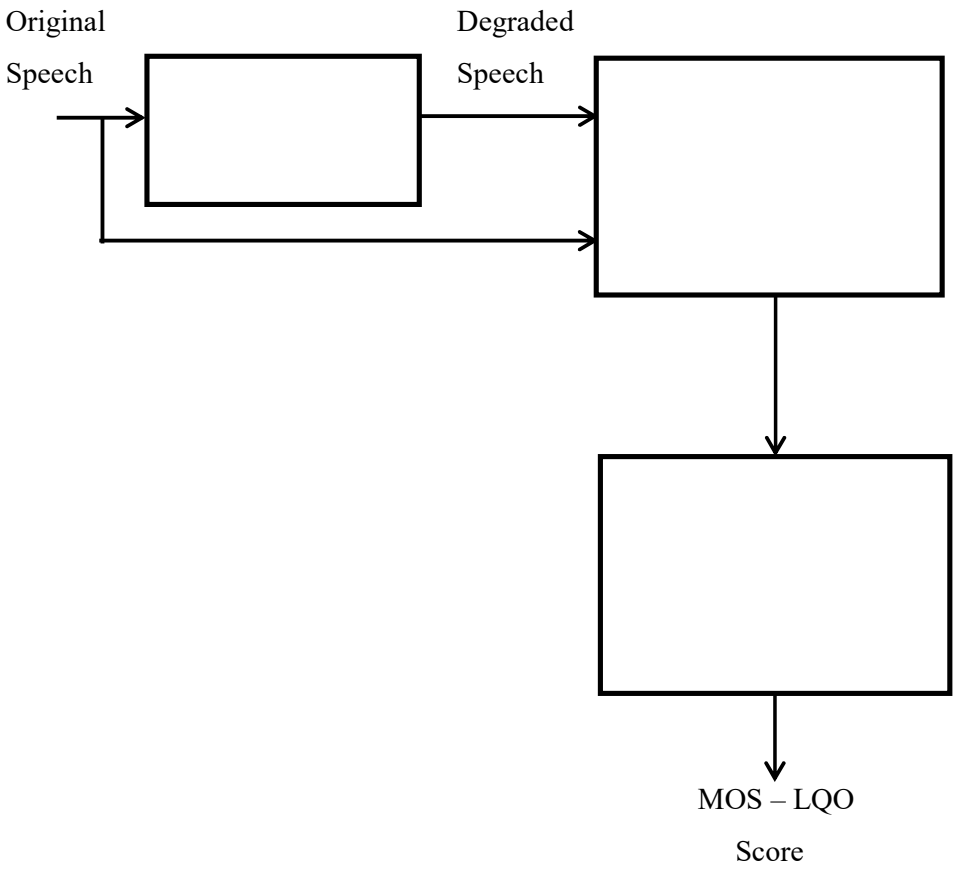



Figure: 2.33. Mapping PESQ Raw Score to the MOS-LQO score.

test (MOS–LQO) in objective assessment of the quality of narrow-band speech (300 – 3400 Hz) transmitted over wireless and other networks, but has poor score coverage.

### *3. The Barriac et al Mapping Function*

Barriac et al in 2004 developed the following mapping function for use with the PESQ algorithm for wideband signals even before the introduction of the wideband ITU-T Rec. P.862.2 mapping function. It is given by:

$$y = 1 + \frac{4}{1 + e^{(-2x+6)}} \tag{2.48}$$

The plot shown in figure 2.35 unfortunately does not correctly cover the range of the raw PESQ score which goes from -0.5 to 4.5.

### *4. The PESQ WB ITU-T P.862.2 mapping function*

The mapping function of PESQ algorithm for wideband speech signals is given by (ITU-T Rec. P.862.2, 2007):

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{(-1.3669x \quad 3.8224)}} \tag{2.49}$$

The function was developed from a set of data from seven subjective experiments made up of five purely wideband speech data sets and two narrowband and mixed speech data sets. It also has poor score coverage of the subjective MOS.

### *5. The Auryst's mapping functions:*

The mapping function developed by Auryst is also a type of logistic function given by Morfitt III and Cotanis (2008):

$$y = a + \frac{b - 1}{1 + e^{c.x+d}} \tag{2.50}$$

where parameters *a*, *b*, *c*, and *d* are coefficients which were optimized for the mapping. Unfortunately, this function ended with only unknown parameters.

### 6. *Morfitt III and Cotanis Logistic Function*

The logistic mapping function developed by Morfitt III and Cotanis (2008), was aimed at achieving improvements in the accuracy of mapping the raw PESQ scale to the subjective MOS scale. It is given by:

$$y = 1 + \frac{4}{1 + e^{(-1.7244x + 5.0187)}} \tag{2.51}$$

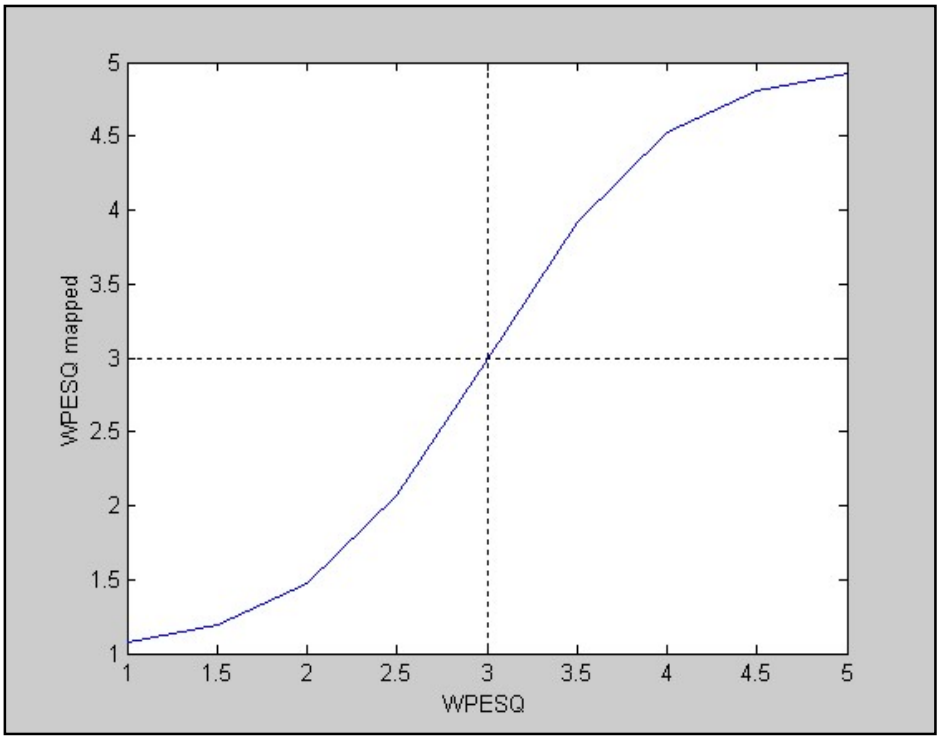



Figure 2.34. The Barriac et al Mapping Function(Barriac et al, 2004).

where,*x* is the raw PESQ Score and *y* is the mapped PESQ MOS-LQO. This function is acclaimed to be more accurate than earlier ones and provided better fit and improves PESQ algorithm performance, but still leaves room for improvements.

## 7. The S-Curve Logistic Model

Most of the mapping functions reviewed above are manipulated versions of the logistic population growth functions. The logistic growth model is a reliable forecast or prediction model for functional changes. The function was originally developed as a differential equation by Verhauslt's in 1838 (Ji, 2013) and given by:

$$\frac{dP}{dt} = r_{max} P\left(1 - \frac{P}{K}\right) \quad (2.52)$$

where, $P$ is the population size that ultimately grows to the carrying capacity, $K$, attime infinity, and $r_{max}$ is the maximum growth rate which occurs at the point of inflection where exponential growth stops and growth or functional change continues as bounded exponential growth.

The function is represented as a simple sigmoidal S-curve with all its important stages shown in Figures 2.36(Kucharavy and Guio, 2015). The carrying capacity, $K$ is a point of saturation or stability of the population, while $\left(1 - \frac{P}{K}\right)$ is the fractional deficiency of the instantaneous population function from the peak, $K$.

Shown in Figure 2.36are the three parameters required to fit the curve, which are: the saturation or carrying capacity, $K$, the growth rate, $r$, and the mid-point or inflection value.The logistic function attains a curve that is similar to the sigmoid curve which onlylies between 0 and 1 in what is known as binary-based logistic regression model. The parameters of the Sigmoid function are $K = 1$, $k = 1$, and $x_0 = 0$, and therefore given by:

$$f(x) = \frac{K}{1 + e^{-k(x - x_0)}} = \frac{1}{1 + e^{-kx}} \quad (2.53)$$

## 2.15 Estimating the Optimal Logistic Parameters

To estimate the parameters of the logistic function, Meng et al (2014) noted that the function can be linearised and the parameters of the linear function obtained. By adopting linear regression the parameters of the logistic function are obtained. In comparing the measured function values with the function values obtained from

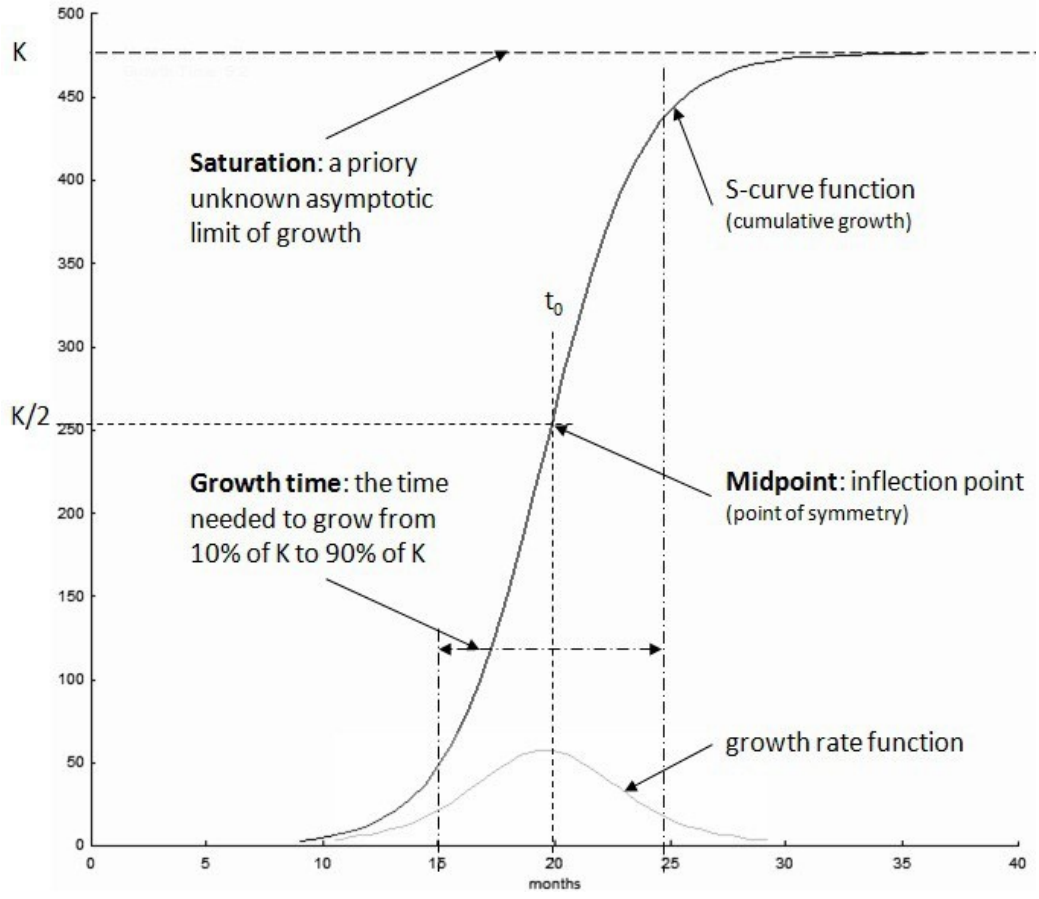



Figure2.35. Schematic diagram of a simple logistic S-curve, defined by three fitting parameters. (Source: Kucharavy and Guio, 2015).

calculations made with substituted logistic parameter values obtained from linear regression, error exist between them. This error is the sum of squares error.

Some other methods discussed in literature for estimation of the parameters of logistic function include the following analytical methods (Skrobacki, 2007):

1. Hotelling's method
2. Tintner's method
3. Bonus's method

Skrobacki noted that these methods were discussed in Stanisz's publication of 1986, which was purely in Polish language, and that the most accurate of them is the Tintner's method. But in his paper, Skrobacki discussed the first two methods.

The Hotelling's method is a type of linearisation which is discussed in literature as making the relative rate of growth of the logistic function as the function of the least square model, given by:

$$\frac{dy/dx}{y} = r\left(1 - \frac{y}{K}\right) \tag{2.54}$$

and linearised as: Y = QX + R, where R = r and Q = $-\frac{r}{K}$.

The Tintner's method entails transforming the function of the logistic model into a new set of functions before linearisation, given by:

$$(z_x, z_{x+1}) = \left(\frac{1}{y_x}, \frac{1}{y_{x+1}}\right) \tag{2.55}$$

where, $z_x = \frac{1}{y_x} = \frac{1+C\text{ex }(-bx)}{K}$, and $z_{x+1} = \frac{1}{y_{x+1}} = \frac{1+C\exp(-b(x+1))}{K}$

Manipulating these equations of $z_x$ and $z_{x+1}$, we have the following equation:

$$z_{x+1} = \exp(-b)\, z_x + \frac{1 - \exp(-b)}{K} \tag{2.56}$$

Equation (2.56) is a linear equation represented by:

$$z_{x+1} = Sz_x + T \tag{2.57}$$

where, $S = \exp(-b)$ and $T = \frac{1-\exp(-b)}{K}$.

Effort at finding the optimal values of the logistic parameters that minimise the error existing in the least squares estimationrequired optimisation to be done. This would move us very close to parameter values that are more accurate and reliable.

Generally speaking, anon-linear optimisation problem is expressed as follows:

Minimise $f(x_1, x_2, \dots, x_n)$ (2.58)

subject to $\phi_i(x_1, x_2, \dots, x_n) = 0, (i = 1, 2, \dots, l)$ (2.59)

$\Psi_j(x_1, x_2, \dots, x_n) \leq 0, (j = 1, 2, \dots, m)$ (2.60)

In vector form we have the function as $f(X)$ where $X = (x_1, x_2, \dots, x_n)$.
The following descriptions of the optimisation problem are with reference to (Soliman and Mantawy, 2012; Chandra et al, 2009; More and Wright, 1993). The optimisation problem might be constrained or unconstrained. For the unconstrained problem, provided the optimisation problem is continuous and differentiable, and for a feasible region, *S*, and a sequence $\{x^k\} = S$ for which $S = \mathbb{R}^n$, we need to solve for the parameter vector $X = [x_1, x_2, \dots, x_n]^T$ that will minimise the objective function.

The partial derivative is obtained as a function of the variables $(x_1, x_2, \dots, x_n)$ and equated to zero respectively. Next, if could be found, the second-order partial derivatives are obtained and also equated to zero to find the Hessian matrix, *H*, of the second-order derivatives. Positive definite *H* implies a minimum point at the solution values of the variables and parameters; otherwiseit implies a maximum point.

Where the function to be optimised is constrained as in (2.58) to (2.60) and $S \subset \mathbb{R}^n$, being equality constrained the Lagrange's multiplier is used to obtain the alternative form of the augmented objective function. Whereas when the function is inequality constrained, the Kuhn-Tucker multiplier is deployed.

Non-linear least squares problems are usually found in data-fitting applications and belong to the unconstrained optimisation family, which has the general form given by:

$min\{f(x) : x \in \mathbb{R}^n\}$ (2.61)

whereby a local minimiser of a real-valued *f* defined on $\mathbb{R}^n$ is sought for, that is, a vector $x^* \in \mathbb{R}^n$ such that $f(x^*) \leq f(x)$ for all $x \in \mathbb{R}^n$ near the minimal point,$x^*$.

Efficient parametric optimisation is built around the Hessian (second–order derivative) approximation based on application of Newton's principle, rather than the gradient descent (first-order derivative) method (Bonnans et al, 2006). The Gauss-

Newton (GN) and the Levenberg-Marquardt (LM) optimisation techniques belong to this (the Hessian) approach.

To calculate error in the optimisation so as to determine the goodness of fit of the model, either Mean Square Error (MSE) or Root Mean Square Error (RMSE) is used is given Villanueva and Ferjoo (2016), Ji (2013), and Li and Jiang(2013):

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(P(t_i) - P_m(i))^2 \tag{2.62}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P(t_i) - P_m(i))^2} \tag{2.63}$$

where, $P_r$ is the rated function, $N$ is the number of pairs of data points, $P(t_i)$ is the function obtained at point $t_i$, $P_m(i)$ is the given function data at point $t_i$.

# CHAPTER THREE
# METHODOLOGY

## 3.1 Introduction

Works carried out to address the problem statement as well as the aim and objectives of this research covered the following areas:

1. Speech acquisition, recording and format conversion.
2. Speech transmission through two intra-networks and one inter-network.
3. Computation of psychoacoustic parameter of loudness for reference and received speeches and development of comparative speech quality measure from results obtained for loudness parameters using a preferred loudness estimation model.
4. Conducted subjective quality measure for the received speeches using Absolute Category Rating (ACR) on listening-only technique.
5. Conducted objective speech quality testing for all received speeches using PESQ model and carried out mapping using an existing internationally standardised mapping function (ITU-T P.862.1).
6. Correlated quality scores from both subjective and objective quality measures.
7. Evaluation of existing mapping functions using raw PESQ scores data obtained from transmission of speeches through a couple of mobile wireless telephone networks was carried out.
8. Development of an improved logistic mapping function to address the constraints of existing measures and techniques, through optimisation of the function parameters using Levenberg-Marquardt optimisation algorithm.
9. Comparative analysis of two international mapping functions and obtained improved mapping function was done using Analysis of Variance (ANOVA).

## 3.2 Speech Acquisition and Processing

### 3.2.1 Speech Recording

To ensure naturalness and intelligibility of raw speech signals used for this work, the speech database used was developed locally. Nonetheless, the following guidelinesin (ITU-T Rec. P.830) for high quality speech recording required for quality testing of received speeches were adhered to in this exercise.

1. Speeches were recorded, collated, sorted and processed in a treated sound recording studio/environment with nominal sound level not above 30dBAinside the studio.
2. Human speakers were required to pronounce words fluently not stylishly, to maintain constant speech level that they found comfortable, and to avoid making noise through any means like rustling of paper, moving of feet on the floor, and so on.
3. Professional recording equipment was used for recording of speeches. The Focusrite Scarlett Studio Pack was purchased and used for speech recording. The pack consists of the Scarlett Solo computer audio interface, CM25 professional condenser microphone, HP60 headphones, Red XLR microphone cable (3 m), Type 'A' – Type 'B' USB cable, Software Activation card with codes for accessing on-line resources which include the driver software, Scarlett Plug-in Suite, CUBASE DAW Software, Loop-Masters sample library, and Multi-language User Guides. The computer-based Focusrite Scarlett Solo Studio interface unit shown in Figure 3.1connects to a computer as shown in Figure 3.2. It has input connections for vocaland guitar recording. It also provides monitoring for playback through headphones or loudspeakers.
4. The CM25 condenser microphone is powered by an inbuilt 48 volts source from the interface, with a light indicator to show the supply is working.
5. The speech recording was done as shown in Figure 3.3 using the CUBASE computer software, which is a multi-track Digital Audio Workstation (DAW) application software shown in Figures 3.4.
6. Four (4) different speech statements were recorded per speaker from the speaker population of eight (8) males and eight (8) females which made a total of 64 original speeches that were recorded.

7. Each speech was segmented into a pair of short sentences read out by each speaker for duration of about three seconds for each half of the pair and the pair was separated with a pulse of two seconds. These made total time duration of eight seconds per complete speech sample.


Figure 3.1. The Focusrite Scarlett Audio Recording Interface Unit.


Figure 3.2. Focusrite Scarlett Speech Recording Setup

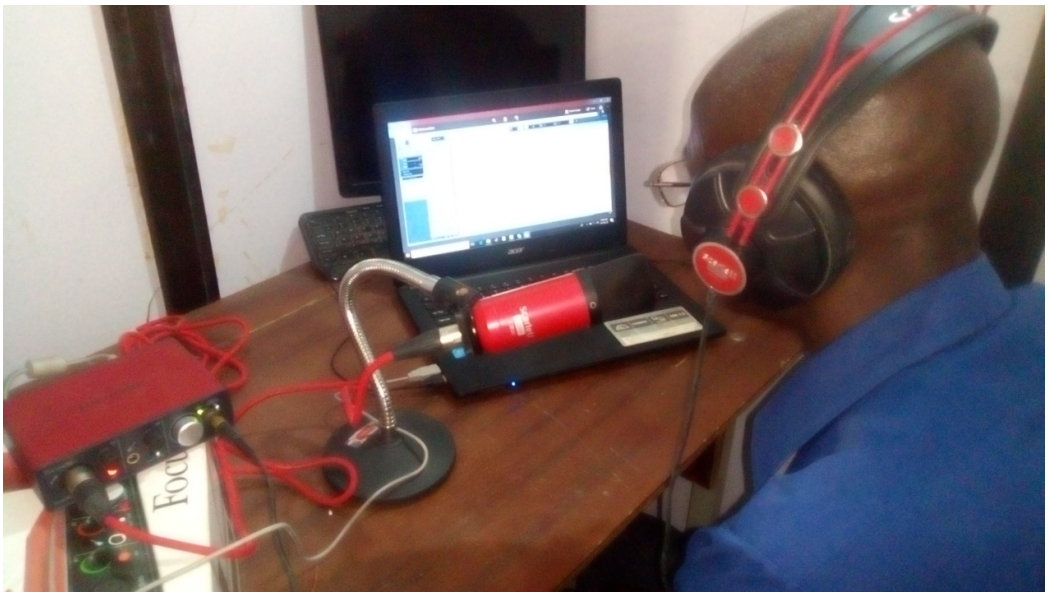


Figure 3.3. Speech Recording Using the Focusrite Scarlett Studio Pack.

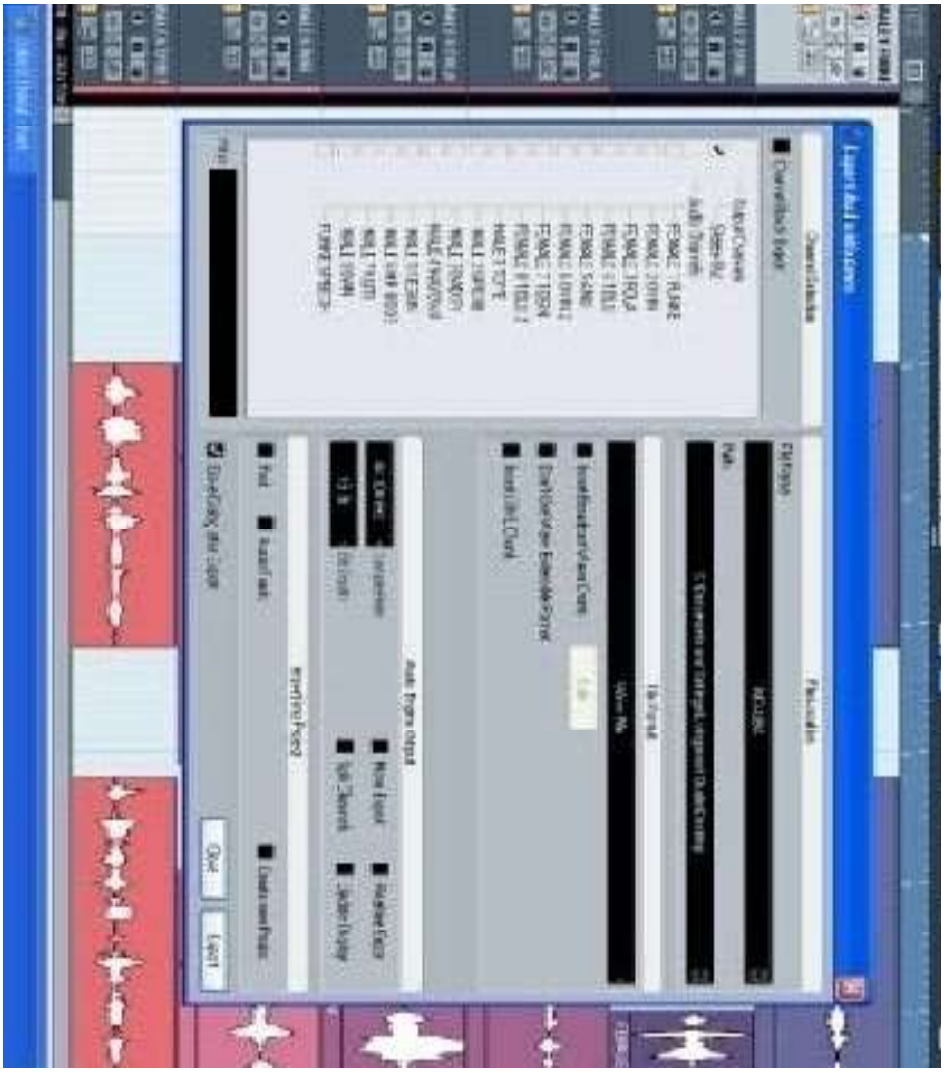


Figure 3.4. CUBASE Software Display DuringRecording of Speeches.

### 3.2.2 Speech Conversion

Speeches recorded into the computer on the CUBASE software were in the Adaptive Multi-Rate (AMR) speech format. The AMR format is a lossy format

which in comparison with some common formats like the MP3, AAC, and WMA formats is especially efficient at compressing and storing voice recordings (Harris, 2018). But, for processing, the speech files were converted from the AMR format to the Waveform Audio (WAV) file format using the 'Any Audio Converter (AAC)' software package.The conversion is necessary for ease of making the speech files storable, processable, editable and transmittable.

The 'wav' audio file format is a Microsoft/IBM standard format for storing raw and uncompressed audio bit-streams on computers. It makes use of a bit-stream Resource Interchange File Format (RIFF) and encoded in Linear Pulse Code Modulation (LPCM)technique, which haslinearly uniform quantization levels.

### 3.2.3 Transmission of Original Speeches

The recorded original speeches were transmitted through some of the existing WCDMA (3G) mobile telecommunication networks in Nigeria. The speeches were transmitted in categories as follows:

**Category 1: Intra-network transmission:**

Network A = original speeches transmitted over Network A = 64 received speeches

Network B = original speeches transmitted over Network B = 64 received speeches

**Category 2: Inter-network transmission**

Network C = original speeches transmitted over an inter–network = 64 received speeches

### 3.2.4 Conversion of Received Speeches

A total of 192 received (degraded) speeches were recorded using a Call Recorder application on the receiving mobile phone, one at a time to build the database of received speeches. These recorded received speeches were in MP3 format, and required conversion to the 'wav' format for processing. A speech file format converter (the Any Audio Converter software) was used to convert the degraded speeches from MP3 to 'wav' file format.

Table 3.1. Table of Legend for Original and Received Speeches.

| **Male** | **Description** | **Female** | **Description** |
|---|---|---|---|
| OMxSy | Original Male x Speech y | OFxSy | Original Female x Speech y |
| AMxSy | Male x Received Speech y over Network A | AFxSy | Female x Received Speech y over Network A |
| BMxSy | Male x Received Speech y over Network B | BFxSy | Female x Received Speech y over Network B |
| CMxSy | Male x Received Speech y over Network C | CFxSy | Female x Received Speech y over Network C |

## 3.3 Quantitative measure of psychoacoustic parameter of speech

In evaluating E2E quality of speech transmitted over the wireless mobile networks, assessment of the key psychoacoustic parameter of speech majorly affected

by distortions and attenuations from network equipment and the transmission channel was carried out. Adopted for this assessmentwas the Zwicker and Fastl non-stationary sound model which wasbuilton principles of the auditory system utilising spectra extraction of speech signals. It was so chosen because Zwicker's parameters
are more appropriate inreflecting physiological processes and perceptual results than the Moore and Glasberg and other available loudness models (Volk, 2016).

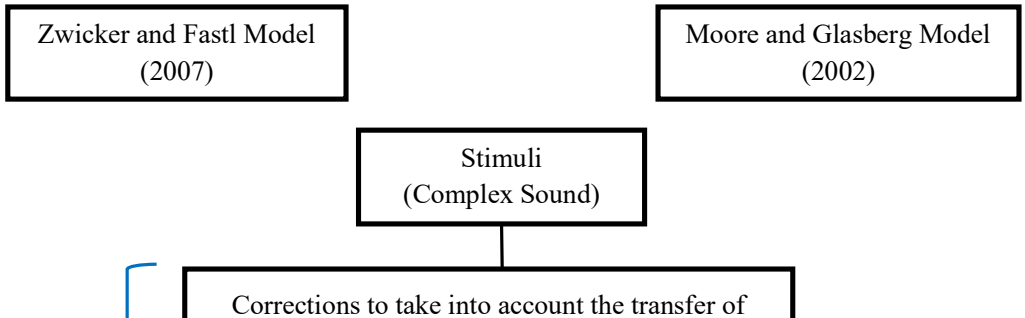
The Zwicker and Fastl model provides computational procedures for estimating loudness and loudness level of sound perceived by listeners with normal hearing condition. It does so by analysing the physical characteristics of sound under given listening conditions.

Processing and computational steps in Zwicker and Fastl model are as follows (Zwicker and Fastl, 2007,Volk, 2016):

1. Sound signal picked-upfrommicrophone orrecorded,
2. Signal amplification,
3. Two-third octave filter-bank filtering,
4. One-way rectification,
5. 2 ms time constantLow-Pass filtering.

Figure 3.5shows a summary of stages for calculating the loudness of time-varying sounds with the use of either the Zwicker and Fastl model or the Moore and Glasberg model.

Statistical indicators for Zwicker and Fastl model useful in finding the global loudness for a particular sound include percentile loudness (Nx – N4, N5, and N7), whichspecify loudness valuesthat were reached during *x* time percentage. For Moore and Glasberg model, to estimate overall perceptual loudness of a time-varying sound,we have to calculate the maximum Short-Term Loudness (STLmax). But to estimate the overall loudness of steady-time or very slowly time-varying sounds we have to calculate the maximum Long-Term Loudness (LTLmax) (Sechadriand Yegnanarayana, 2009).



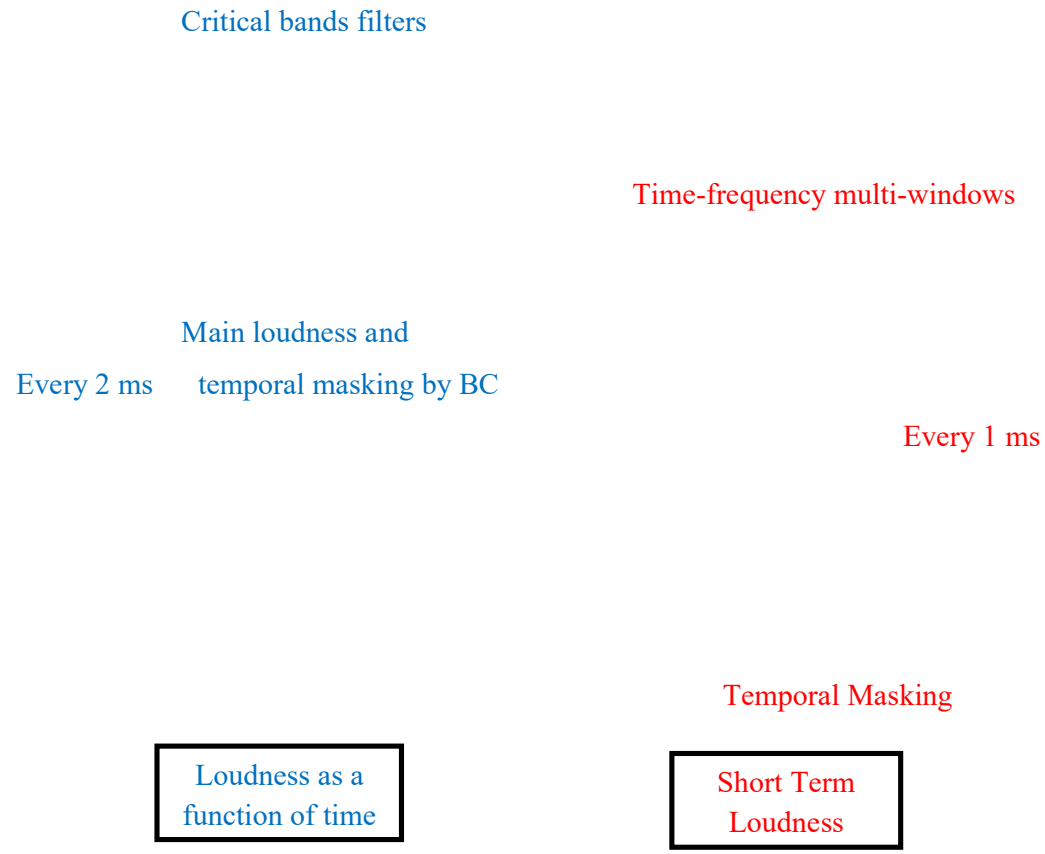



Figure 3.5. CalculationStepsforLoudness ofTime-Varying Sounds. (Source: Genesis IB/RP/10003, 2009).

***Key:***

Activities typed in blue colour belong to the Zwicker and Fastl Model.

Activities typed in red colour belong to theMoore and Glasberg Model

Activities typed in black are common to the two models.

Loudness calculation using Zwicker and Fastl model entailed:

1. Obtain filter transfer that represents outer to middle ear functions;

2. Calculateexcitations, *E*, by making use of auditory filter banks to represent inner ear functions, notingvery important roles played by critical bandwidth and frequency selectivity of the auditory system in loudness.
3. Utilise power law relationsto transform excitation patterns into specific loudness, *N'*.

Specific loudness was calculated using the following equation(Zwicker and Fastl, 2007, Webster and Jiricek, 2014):

$$N' = \alpha\left(\frac{E_{THQ}}{E_0}\right)^{\beta}\left[\left(0.5 + 0.5\frac{E}{E_{THQ}}\right)^{\beta} - 1\right]\frac{\text{sone}_{\text{G}}}{\text{Bark}} \tag{3.1}$$

where: $\alpha$ and $\beta$ are constants given by $\alpha = 0.08$, $\beta = 0.23$,$E_{THQ}$ is excitationat threshold in quietness, $E_0$ is excitation with respect to reference intensity, $E$ is the excitation at a specific frequency, while $I_0 = 10^{-12}\ W/m^2$. Soneappendage of letter G implies determining specific loudness with respect to the critical-band levels or rates.

The total loudness is integration of the specific loudness over frequencies in Barks scale, obtained by summing neural operations of sound across BM in the inner ear.

Total loudness therefore was obtained from:

$$N = \int_0^{24Barks} N'dz \tag{3.2}$$

where $N'$ is the specific loudness given in Equation 3.1.

### 3.3.1 Programming of Loudness Estimation

Computation of stages involved inobtaining loudness of time-varying sounds is complex, and in all cases computer software programmes are used to reduce computational stress, time and resources utilization.

Professional sound quality computer software used for calculation of loudness parameter values in this work was the software by Genesis (2009) written in Matlab. The Genesis loudness toolbox implements the loudness algorithms for steady, time-varying and impulsive sounds and it is validated using Matlab software as enumerated on Table 3.2.

Table 3.2. Loudness Models implemented in Loudness Toolbox.

| Matlab function name | Model/ standard | Stationary sound | Time-varying sound | Impulsive sound |
|---|---|---|---|---|
| Loudness_ISO532B | ISO 532<br><br>DIN 45631 | √ | | |
| Loudness_ANSI_S34_2007 | ANSI S3.4-2007 | √ | | |
| Loudness_NonStationnary_ Zwicker | Fastl and Zwicker loudness model for time-varying sounds | | √ | |
| Loudness_NonStationnary_ Moore | Moore and Glasberg loudness model for time-varying sounds | | √ | |
| Loudness_LMIS | Impulse sounds loudness model by Boullet et al | | | √ |

## 3.4 Subjective Speech Quality Testing

The opinion scoresof subjects (listeners) of the quality of the speeches received from transmission over the networks, now being network-degraded were obtained using the Listening–Only Test (LOT) method.This has been directly applied in assessing unidirectional transmission systems like broadcasting, public address, recorded announcement systems, and telephone talks in one direction. Received speeches were played out to subjects who listened to them and indicated their opinions of the quality of the speeches on the ACR MOS rating scale ON Table 3.3.

Total number of received speeches that were rated = 192 speeches

Number of Subjects selected = 20 persons

Subjective test environment = conference format, in a quiet room free from interferences with noise level not over 30dB.

Speech relaying devices = two professional loudspeakers used to play out received speeches for subjects to listen to and to score the quality of the received speech on the scale of 1 to 5.

Mean Opinion Score (MOS): was calculated as the average of opinion scores supplied by subjects for each received speech as follows:

$$MOS = \frac{1}{M}\sum_{i=1}^{M} OS_i \tag{3.3}$$

where,$M$ = 20 is the total number of subjects that participated, and$OS_i$ is the Opinion Score of individual subject for a particular received speech.

## 3.5 Objective Quality Assessment of Received Speeches

The objective perceptual evaluation was carried out for all received speeches using PESQ algorithm adopted for this part of the work. PESQ is a robust first level intrusivemodel for E2Eassessment of speech quality. It was deployed based on the provisions of ITU-T P.862 (02/2001). An overview of the PESQ algorithm is given as follows.

### 3.5.1 Main stages of PESQ algorithm:

1. Level and time alignment
2. Perceptual modeling
3. Determination of the PESQ quality score.

Table 3.3. Subjects Rating Table.

| **Speech No. 001** | | |
|---|---|---|
| Listening quality | Opinion Score | Tick |
| Excellent | 5 | |
| Good | 4 | |
| Fair | 3 | |
| Poor | 2 | |
| Bad | 1 | |

***Instruction:*** Subjects ticked their opinion for each speech listened to from ` Speech 001 to Speech 192 as indicated on the table.

The PESQ algorithm processing steps are summarised as follows:

**Level Alignment:**Level alignment ofthe original signal, $X(t)$and degraded signal, $Y(t)$to the same constant power level, entails the following activities:

Compute the filtered versions of both speech signals and the average value of the squared filtered speech samples.

Calculate different gains and apply them to align both $X(t)$and $Y(t)$to a constant target level. This results in the scaled versions $X_S(t)$and $Y_S(t)$of these signals.

**IRS Filtering:**Intermediate Reference System (IRS) filtered versions of both speech signals are computed to model the signals that the subjects listened to. The steps are:Fast Fourier Transform (FFT) of the file, filtering, and thenan inverse FFT operation. This results in the filtered versions, $X_{IRSS}(t)$and $Y_{IRSS}(t)$of the scaled input and output signals $X_S(t)$and $Y_S(t)$. The IRS filtered signals are required for time alignment procedure and perceptual modeling.

**Time alignment:**Envelopes $X_{\mathrm{ES}}(t)k$ and $Y_{\mathrm{ES}}(t)k$ are calculated from the scaled original and degraded signals $X_S(t)$and $Y_S(t)$, and defined as:

$$LOG\ (MAX(E(k)/E_{thre}\ \ ,I) \tag{3.4}$$

where $E(k)$is the energy in 4 ms frame $k$ and $E_{thres}$ is the threshold of speech determined by a VAD. Cross-correlation of the envelopes for the original and degraded signals is utilized for finding the crude delay between them, with an approximate resolution of 4 ms.

**Fine time alignment:**Perceptual models are sensitive to offsets in time, therefore accurate delay values are calculated for the speech samples as follows:

Original and degraded signals are split into 64 ms frames (75 % overlapping) multiplied with Hann window functions and cross-correlated. A measure of the confidence of the alignment in each frame is obtained to give the delay estimate for each frame. A histogram of these delay estimates is calculated and smoothed by convolution with a symmetric triangular kernel. The index of the maximum in the histogram, combined with the previous delay estimate, gives the final delay estimate.

**Utterance splitting:**Repeated splitting and realigning time intervals in each utterance is done to test for delay changes in speech until the greatest confidence is identified.

**Perceptual Modeling:**This calculates a distance between the original and degraded speech signals, known as PESQ score, using a short-term FFT with a Hann window of size 32 ms. With this the time signals are mapped to the time-frequency domain.

The values at the center of the Bark bands are obtained from interpolating the absolute hearing threshold, $P_0(f)$. These values are stored in an array and are used in Zwicker's loudness formula to obtain a warped loudness scale.

The power representations $PX_{WIRSS}(f)_n$and $PY_{WIRSS}(f)_n$ and the pitch power densities $PPX_{WIRSS}(f)_n$ and $PPX_{WIRSS}(f)_n$ are obtained and the power spectrum of the original and degraded signals averaged over time.

Partial compensation for filtering and short-term gain variations are carried out for both the original and degraded signals, then the pitch power densities are transformed to a Sone loudness scale using Zwicker's loudness law:

$$LX(f)_n = S_l \cdot \left(\frac{P_0(f)}{0.5}\right)^{\gamma} \cdot \left[\left(0.5 + 0.5 \cdot \frac{PPX'_{WIRSS}(f)_n}{P_0(f)}\right)^{\gamma} - 1\right] \tag{3.5}$$

where $P_0(f)$is the absolute threshold, $S_l$ isthe loudness scaling factor and the Zwicker power factor, $\gamma$ , is 0.23. The resulting two-dimensional arrays $LX(f)_n$and $LY(f)_n$are known as the loudness densities.

A signed difference between the loudness density of the original and degraded signals is calculated. A disturbance density as a function of time (window number $n$) and frequency, $D(f)_n$ is also calculated. This is multiplied by an asymmetry factor to obtain an asymmetrical disturbance density $DA(f)_n$for every frame.

The disturbance density $D(f)_n$and asymmetrical disturbance density $DA(f)_n$are integrated (summed) along the frequency axis using two different $Lp$ norms and a weighting on soft frames (having low loudness):

$$D_n = M_n \cdot \sqrt[3]{\sum_{f=1,\ldots Number\ of\ Bark\ bands} \left(|D(f)_n| W_f\right)^3} \tag{3.6}$$

$$DA_n = M_n \cdot \sum_{f=1,\ldots Number\ of\ Bark\ bands} \left(|D(f)_n| W_f\right) \tag{3.7}$$

The aggregated values are called frame disturbances, $D_n$and $DA_n$. At moments when the delay between the signals is negligible, the frame disturbances become $D'_n$and $DA'_n$.

Frames that have frame disturbance above a particular threshold are known as bad intervals for which a new delay value is estimated. For this a new frame disturbance is recomputed, and with lesser disturbance value, the final frame disturbances $D''_n$and $DA''_n$that are used to obtain the perceived quality are obtained.

Finally, the PESQ score is obtained within the range –0.5 to 4.5 as a linear combination of the average disturbance value and the average asymmetrical disturbance value.

### 3.5.2 Programming the PESQ Algorithm

Perceptual Evaluation of Speech Quality (PESQ) testing was conducted using the ITU-T PESQ Source Code which was programmed and compiled withDev C++ compiler.

PESQ code, was obtained from the ITU-T website http://www.itu.int/rec/T-REC-P.862-200511-I!Amd2/enand compiled. To compile the C-code and run the PESQ file,simulated PESQ program was accessed using Command prompt (cmd) launched as a black terminal window shown in Figure 3.6.

Files of original speech, degraded speech and simulated PESQ program were saved in a different folder, which were accessed for each speech file with".wav" extension added to the filenames and sampled at the rate of +8000. This step was repeated for each original speech and the corresponding degraded speech to give PESQ MOS score for the particular speech sample.

## 3.6 Functions for Mapping PESQ Raw Scores

### 3.6.1 Evaluating Existing PESQ Mapping Functions

Before correlating the raw PESQ score obtained from objective testing with the subjective MOS, there is need to first map the raw PESQ scores on the scale of -0.5 to 4.5 to the standard MOS scale of 1 (for bad quality) to 5 (for excellent quality). Two known international standard mapping functions were evaluated for this purpose.

First was the ITU-T Rec. P.862.1 Amendment to PESQ (ITU-T Rec. P.862.1, 2003) given by:

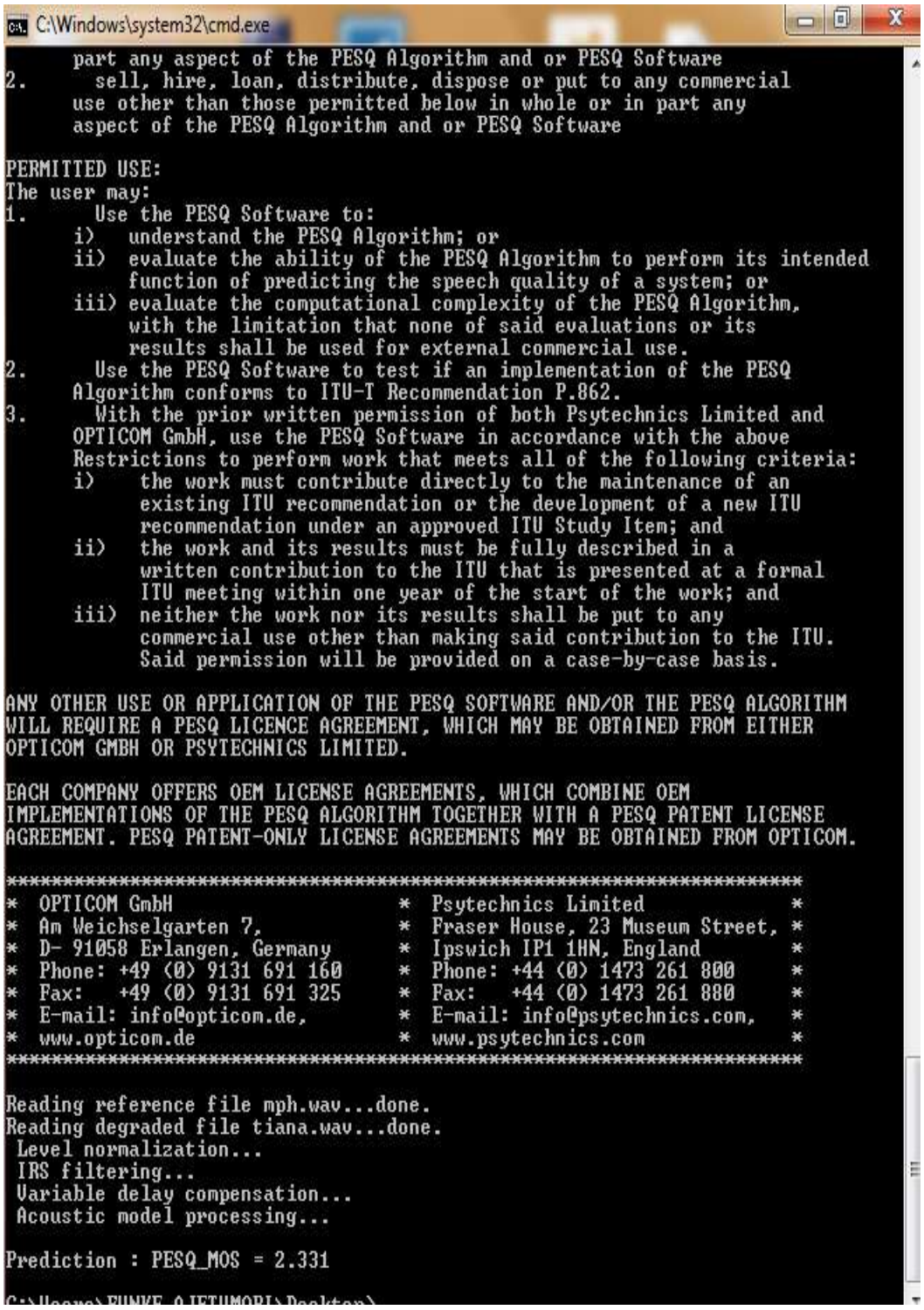



Figure 3.6. Command Prompt Showing PESQ Result.

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{(-1.4945x + 4.6607)}} \quad (3.8)$$

where, $x$ is the raw PESQ score and $y$ is the mapped PESQ score given as PESQ MOS-LQO.

Second was the mapping function invented by (Morfitt III and Cotanis, 2008) patented by the United States, given by:

$$y = 1 + \frac{4}{1 + e^{(-1.7244x + 5.0187)}} \quad (3.9)$$

where$x$ is also the raw PESQ Score and $y$is also the mapped PESQ MOS score (MOS-LQO).

### 3.6.2 Developing ImprovedLogistic Mapping Function

The logistic mapping function was studied and built upon the theories of logistic population growth differential function. Solution was obtained for the logistic growth differential equation (2.52) by partial integration of the equation:

$$\frac{dP}{P\left(1 - \frac{P}{K}\right)} = rdt \quad (3.10)$$

So that,

$$\frac{dP}{P\left(1 - \frac{P}{K}\right)} = \left[\frac{1}{P} + \frac{1}{(K - P)}\right] dP = rdt \quad (3.11)$$

Integrating (3.11) we have:

$$\ln P - \ln(K - P) = rt + c \quad (3.12)$$

$$\frac{P}{K - P} = e^{rt + c} = e^{rt + c} \quad (3.13)$$

Separating terms, the growth function is obtained as:

$$P = \frac{Ke^{rt + c}}{1 + e^{rt + c}} = \frac{K}{1 + e^{-(rt + c)}} = \frac{K}{1 + Ce^{-rt}} \quad (3.14)$$

where, $C = e^{-c}$ is a coefficient obtained at the initial condition(at t = 0), given by:

$$C = \frac{K}{P_0} - 1 \; or \; \frac{K - P_0}{P_0} \quad (3.15)$$

Three key features of logistic growth function given by (Tsoularis and Wallace, 2002), which were also proved, state that:

1. Population size, $P$(t) will eventually reach the carrying capacity, $K$, asymptotically, expressed by:

$$\lim_{t \to \infty} P(t) = K \quad (3.16)$$

2. The relative growth rate, $\frac{1}{P}\frac{dP}{dt}$, declines linearly with increasing population size, and

3. The population at the point of inflection (PI), where growth rate is maximum, is exactly half of the carrying capacity, that is, $P_i = \frac{K}{2}$. This was obtained from the second order derivative of the function (Safuan et al, 2013).

Rewriting the population function as *y*(x), (3.14) becomes:

$$y(x) = \frac{K}{1 + Ce^{-rx}} \tag{3.17}$$

A four-parameter approach consisting of coefficients: *a*, *b*, *c* and *d*, was adopted for full description of the range of steepness of the S-curve and the offsets on *x* and *y* axes of the logistic function, as shown in Figure 3.7. This becomes particularly important because none of MOS scale or raw PESQ scores starts from the zero point.

Parameter *a* is the full range of the growth function while*d*is offset from originon the function axis or the minimum value of the function. Parameter *b*, which is the same as *r* in (3.14), determines the steepness of the curve, and parameter *c*is a factor of the initial value of the function.

With the raw PESQ range between -0.5 and 4.5 on the *x*–axis and the Subjective MOS range between 1.0 and 5.0 on the function axis, offsets on*x* and *y* are such that the initial condition of the logistic function, $y(x_{-0})$, which is not necessarily the same as $y(x_0)$ because of the offset, necessitates that the function be redefined as:

$$y(x) = y(x_{-0}) + \frac{K - y(x_{-0})}{1 + Ce^{-bx}} \quad or \; d + \frac{a}{1 + e^{-bx+c}} \tag{3.18}$$

In Figure 3.7, the offset on the *y*–axis is$y(x_{-0}) = d = 1$. On the *x*–axis, the raw PESQ scale has an offset of $x_{-0} = -0.5$.Actual carrying capacity $K = a + d = 5$, and the point of inflection (PI) obtained from the second derivative of the solution, $y(x)$, is given by:

$$\left[\frac{\ln C}{r},\ \frac{a}{2} + d\right] = \left[-\frac{c}{b},\ 3\right].$$

The function in (3.17) therefore becomes:

$$y(x) = 1 + \frac{4}{1 + Ce^{-b}} = 1 + \frac{4}{1 + e^{-(bx+c)}} \tag{3.19}$$

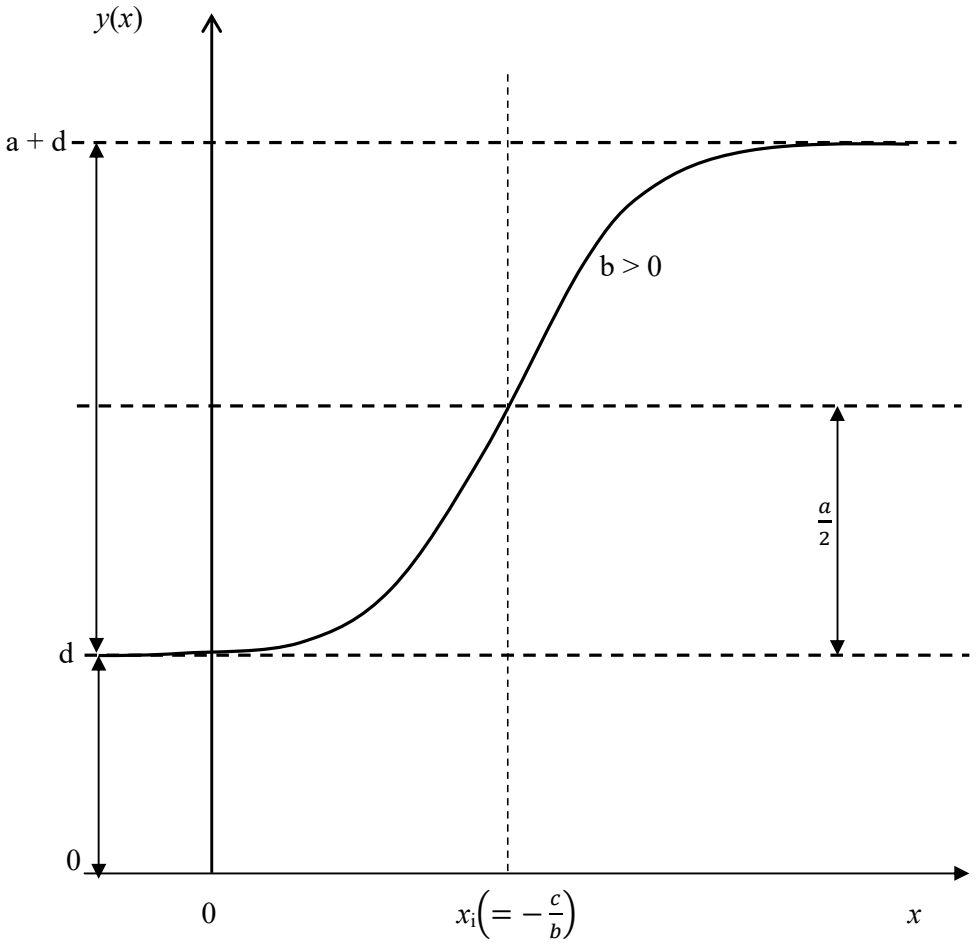



Figure 3.7. Logistic Growth Function with Offset Parameters.

## 3.7 Determination and Optimisation of Logistic Parameters

Parameter extraction or determination and optimisation have as its main goal the determination of model parameters that would minimise differences between measured and estimated model values.From Section 3.6, parameter *a*= 4, parameter *d* = 1; while parameter *b*, the intrinsic rate or steepness of the function curve and parameter *c*, the coefficient of integration determined by the initial value of the function as stated by (3.15), are unknowns. The unknown parameters most be determined and optimised through fitting the logistic function to measured data.

Variations in parameters *b* and *c*, has the capability to bring about improvements in the logistic mapping function which could be achieved through optimizing these parameters. A couple of mathematical techniques were applied for estimating, and optimising parameters *b* and *c*.

### 3.7.1 The Acceleration Function Method

This method was adopted from Mishan et al (2011) with a couple of assumptions made as part of efforts at obtaining the parameters. The first to the third order differentiations of the logistic function (3.17) were obtained as follows:

$$y(x) = 1 + \frac{4}{1 + e^{-(bx+c)}} \equiv 1 + \frac{a}{1 + e^{-(bx+c)}} \tag{3.20}$$

$$\frac{dy}{dx} = \frac{abe^{-bx-c}}{(1 + e^{-bx-c})^2} \tag{3.21}$$

$$\frac{d^2y}{dx^2} = \frac{ab^2e^{-bx-c}(e^{-bx-c} - 1)}{(1 + e^{-bx-c})^3} \tag{3.22}$$

$$\frac{d^3y}{dx^3} = \frac{ab^2e^{-bx-c}[(e^{-bx-c})^2 - 4e^{-bx-c} + 1]}{(1 + e^{-bx-})^4} \tag{3.23}$$

The differentials were equated to zero to obtain solution coordinates of useful parameters or points on the function shown in Figure 3.8.

$$\frac{d^2y}{dx^2} = 0$$

produced: $x = -\frac{c}{b}, y = \frac{a}{2}$

This is the Acceleration Growth Function (AGF), and it produced the maximum point of the growth rate (*dy/dx*) which is the Point of Inflection (PI) of the function *y*.

$$\frac{d^3y}{dx^3} = 0$$

produced: $x_1 = -\frac{[\ln(2+\sqrt{3})+c]}{b} =, y_1 = \frac{a(3-\sqrt{3})}{6}$ ; and

$x_2 = -\frac{[\ln(2-\sqrt{3})+c]}{b}, y_2 = \frac{a(3+\sqrt{3})}{6}$

The Maximum Acceleration Point (MAP): this is the point where the function attains the sharpest rate increase. The coordinates of which are the first solution obtained from the third derivative: $\left(x_1 = -\frac{[\ln(2+\sqrt{3})+c]}{b}, y_1 = \frac{a(3-\sqrt{3})}{6}\right)$.

The Maximum Deceleration Point (MDP): from the PI deceleration begins up to where the deceleration is maximum or the acceleration is minimum; the coordinates of which are the second solution ($x_2$, $y_2$) from the third derivative:

$\left(x_2 = -\frac{[\ln(2-\sqrt{3})+c]}{b}, y_2 = \frac{a(3+\sqrt{3})}{6}\right)$.

### 3.7.2 Non-linear Least Squares Regression Problem

The logistic equation (3.19) is a non-linear equation which cannot be solved analytically. It is therefore classified as a non-linear least squares problem (Jukic and Scitovski, 2003, Matthew, 1992), and the method of data linearisation was used to reduce it to finding the values of the parameter of a least squares line to obtain initial values for the function parameters.

Rearranging the very logistic part of (3.19), such that:

$$y(x) = 1 + \frac{4}{1+Ce^{-bx}} = 1 + g(x) \tag{3.24}$$

We have:

$$\ln\left(\frac{4}{g(x)} - 1\right) = \ln C - bx \tag{3.25}$$

Changing variables by taking $Y = \ln\left(\frac{4}{g(x)} - 1\right)$, $K = \ln C$, $B = -b$, and $X = x$, we have the following linear form:

$$Y = BX + K \tag{3.26}$$

Data points used for plotting this linear function are the transformed (linearised) data of the data obtained from experimental work, such that:

$$(X_i, Y_i) = \left[x_i, \ln\left(\frac{4}{g(x)} - 1\right)\right] \tag{3.27}$$

for $i = 0, 1, 2, \ldots, N$. $X_i$ is chosen as $X_i = x_i - x_0$.

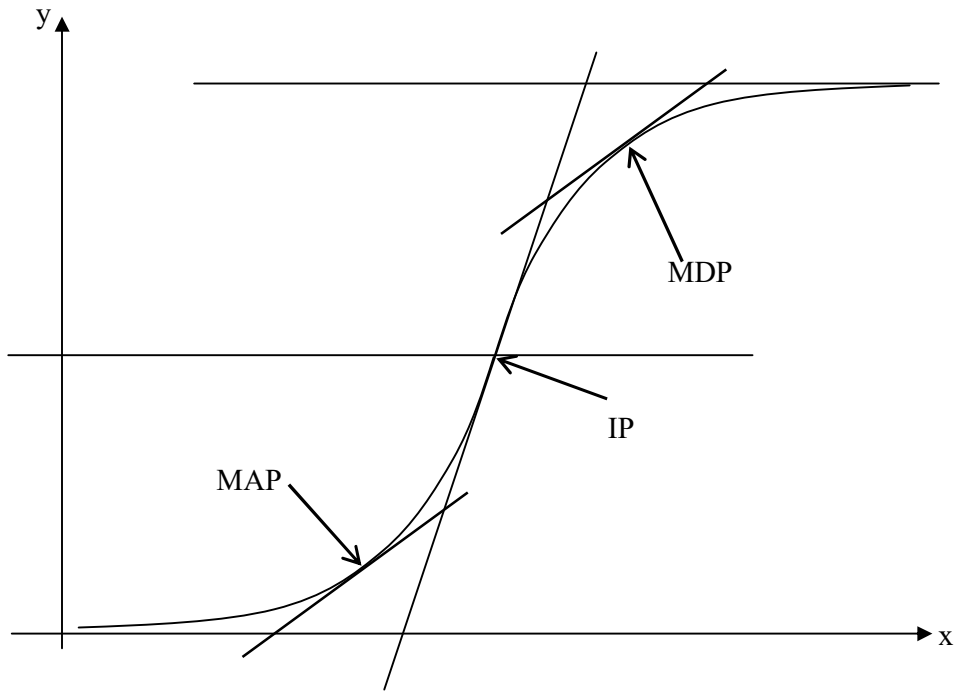



Figure3.8. Logistic Curve indicating Critical Points of the Acceleration function.

Transformed data points, $(X_i, Y_i)$, were plotted using MATLAB. Coefficients B and K were obtained from fitting the least squares regression line to the transformed data. Then the coefficients of the logistic function were obtained as: $C = e^K$, and $b = -B$, from the equation of the linear regression obtained from the MATLAB plot.

### 3.7.3 The Levenberg-Marquardt Optimisation Technique

The Levenberg-Marquardt (LM) algorithm was originally developed by Levenberg in 1944 and reinvented by Marquardt in 1963. It has experienced a couple of improvements and variants of it have been developed over the years. It is an unconstrained iterative optimisation technique used for approximating or extracting parameters of non-linear functions ora set of non-linear equations. It is unconstrained since no conditions were imposed on the independent variable(s) and the function is defined for the range of all values of the independent variable(s).

The LM algorithm is a hybrid techniquethat combines the features of the gradient (steepest) descent and the Gauss-Newton (GN) algorithms to find the minimum of a function in a non-linear least squares problem. When the process of obtaining a solution is far from a local minimum, the algorithm is slow and behaves like gradient descent, butbehaves like GN when close to the minimum and at such point converges very fast (Levenberg, 1944; Marquardt, 1963; Duc-Hung et al, 2012). It therefore possesses main advantages of both algorithms: stability for GradientDescent (GD) and speed for GN. But, it is more robust than either of them, in that it locates the local minimum faster even when the starting point is far from it.

While the GD algorithm is a first-order expansion of the Taylor series of a non-linear function,the GN algorithm is a second-order expansion of the Taylor's series and difficult to calculate because of the Hessian matrix that needed to be calculated. Advanced algorithms that were developed to overcome this difficulty approximate the Hessian matrix, and these include the Leverberg-Marquardt algorithm (Sarabakha et al, 2017).

For a set of observed data, $y_i$, best-fit parameters minimise the objective function for non-linear least squares problems given byRanganathan (2004); Madsen et al (2004):

$$\Phi(x_i, \mathbf{p}) = \frac{1}{2}\sum_{i=1}^{m}[y_i - y(x_i, \mathbf{p})]^2 = \frac{1}{2}\sum_{i=1}^{m}[\boldsymbol{r}_i(x)]^2 \qquad (3.28)$$

In vector form,

$$\Phi(x_i, \mathbf{p}) = \frac{1}{2}\|\boldsymbol{r}(x)\|^2 \quad (2.29)$$

where,$i$ = 1, 2, …, $m$;$m$ is number of data points;$x$ is a vector of independent data points;$\mathbf{p}$ is parameter vector such that $\mathbf{p} \in \mathbb{R}^n$;$\boldsymbol{r}_i$ is a residual vector such that $\mathbb{R}^n \rightarrow \mathbb{R}$, while $n$ is the number of parameters in the non-linear function.

The LM algorithm evolved from modifications carried out on the GN algorithmby the introduction of a damping factor which made it to be referred to as a damped least-square technique (More and Wright, 1993). This made for its effectiveness and popularityin solving non-linear least squares problems. Despite the effectiveness of LM at converging fast to the local minimum with few iterations, its accuracy could be affected by the initial and true value. Also, the fact of its being highly computational, particularly in matrix inversions, limits it to applications where the number of parameters to be optimised is not very large (Zhang et al, 2013; Ouadfeul and Aliouane, 2015).

Description of the methodology and procedural concepts of LM algorithm and its implementation found in literature are summarised below (Lourakis, 2005; Duc-Hung et al, 2012; Wikipedia, 2018; Gavin, 2019).

The optimisation process is started with a guessed value and iterated towards the optimal value. Starting at the initial value of a parameter, $\mathbf{p}_0$, determined by operation of least squares regression initially carried out, the last value of $\mathbf{p}$ is updated with addition of a $\delta_{\mathbf{p}}$. Operation of LM produces a series of parameter vectors: $\mathbf{p}_1$, $\mathbf{p}_2$, $\mathbf{p}_3$, …, $\mathbf{p}_n$, that converge at a local minimum $\mathbf{p}^+$ for the function, where $n$ is the number of parameters.

At each iteration the task is to find $\delta_{\mathbf{p}}$ that will minimise the following, which took a clue from Taylor series expansion:

$$||x - f(\mathbf{p} + \delta_{\mathbf{p}})|| \approx ||x - f(\mathbf{p}) - \mathbf{J}\delta_{\mathbf{p}}|| = ||\boldsymbol{\epsilon} - \mathbf{J}\delta_{\mathbf{p}}|| \quad (3.31)$$

where,$\boldsymbol{\epsilon}$ is residual error vector, that is, differences between measured and estimated values of the function, given by:

$$\boldsymbol{\epsilon} = \begin{Bmatrix} e_{11} \\ e_{12} \\ \cdots \\ e_{1m} \\ \cdots \\ e_{p1} \\ e_{p2} \\ \cdots \\ e_{pm} \end{Bmatrix} \tag{3.32}$$

**J**,the Jacobian matrix is a $m \times n$ matrix of partial derivatives of the errors with respect to the parameters,$\frac{\partial f(x_i,\mathbf{p})}{\partial \mathbf{p}}$.It is a column space,introduced to simplify the calculation, and given by:

$$\mathrm{J} = \begin{Bmatrix} \frac{\partial f(x_1,\mathbf{p_1})}{\partial \mathbf{p_1}} & \frac{\partial f(x_1,\mathbf{p_2})}{\partial \mathbf{p_2}} & \cdots\cdots\cdots & \frac{\partial f(x_1,\mathbf{p_n})}{\partial \mathbf{p_n}} \\ \frac{\partial f(x_2,\mathbf{p_1})}{\partial \mathbf{p_1}} & \frac{\partial f(x_2,\mathbf{p_2})}{\partial \mathbf{p_2}} & \cdots\cdots\cdots & \frac{\partial f(x_2,\mathbf{p_n})}{\partial \mathbf{p_n}} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(x_m,\mathbf{p_1})}{\partial \mathbf{p_1}} & \frac{\partial f(x_m,\mathbf{p_2})}{\partial \mathbf{p_2}} & \cdots\cdots\cdots & \frac{\partial f(x_m,\mathbf{p_n})}{\partial \mathbf{p_n}} \end{Bmatrix} \tag{3.33}$$

The Jacobian can be calculated by either numerical approximation using numerical finite difference and secant approximation or analytic methods.The analytic calculation of Jacobian when is possible, makes the iteration to converge faster (Comandur, 2011).

Finding $\delta_{\mathbf{p}}$ is done by linearising the function with updated parameter value, that is:

$$f(x_i, \mathbf{p} + \delta_{\mathbf{p}}) \approx f(x_i,\mathbf{p}) + \mathbf{J}\delta_{\mathbf{p}} \tag{3.34}$$

Therefore, rewriting (3.31), we have:

$$\Phi(x_i, \mathbf{p} + \delta_{\mathbf{p}}) \approx \frac{1}{2}\sum_{i=1}^{m}\left[y_i - f(xi, \mathbf{p}) - \mathbf{J}\delta_{\mathbf{p}}\right]^2 \tag{3.35}$$

In vector form: $\Phi(x_i, \mathbf{p} + \delta_{\mathbf{p}}) \approx ||y_i - f(\mathbf{p}) - \mathbf{J}\delta_{\mathbf{p}}||^2$

$$= [y_i - f(\mathbf{p}) - \mathbf{J}\delta_{\mathbf{p}}]^{\mathrm{T}}[y_i - f(\mathbf{p}) - \mathbf{J}\delta_{\mathbf{p}}]$$

$$= [y_i - f(\mathbf{p})]^{\mathrm{T}}[y_i - f(\mathbf{p})] - [y_i - f(\mathbf{p})]^{\mathrm{T}}\mathbf{J}\delta_{\mathbf{p}} - (\mathbf{J}\delta_{\mathbf{p}})^{\mathrm{T}}[y_i - f(\mathbf{p})] + \delta_{\mathbf{p}}{}^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathrm{J}\delta_{\mathbf{p}}$$

$$= [y_i - f(\mathbf{p})]^{\mathrm{T}}[y_i - f(\mathbf{p})] - 2[y_i - f(\mathbf{p})]^{\mathrm{T}}\mathbf{J}\delta_{\mathbf{p}} + \delta_{\mathbf{p}}{}^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{J}\delta_{\mathbf{p}} \tag{3.36}$$

Differentiating $\Phi(x_i, \mathbf{p} + \delta_{\mathbf{p}})$ with respect to $\delta_{\mathbf{p}}$ and equating to zero gives:

$$(\mathbf{J}^{\mathrm{T}}\mathbf{J})\delta_{\mathbf{p}} = \mathbf{J}^{\mathrm{T}}\,[y_i - f(\mathbf{p})] = \mathbf{J}^{\mathrm{T}}\epsilon \qquad (3.37)$$

whereby $\mathbf{J}^{\mathrm{T}}\mathbf{J}$is Hessian approximation to the matrix of the second-order derivatives. given by: $\nabla^2 f(x) = \left(\partial_j \partial_i f(x)\right)$.

Equation (3.37)is the equation of the GN algorithm, while LM algorithm is achieved as an improvement on it by a slight adjustment brought about by the introduction of a damping factor, λ, which led to "augmented normal equations" (Lourakis, 2005) given by:

$$(\mathbf{J}^{\mathrm{T}}\mathbf{J} + \lambda\mathbf{I})\delta_{\mathbf{p}} = \mathbf{J}^{\mathrm{T}}\epsilon \qquad (3.38)$$

The damping factor, λ (a non-negative factor), is used to update the iteration process. It is adjusted during each iteration stepuntil the sum of squares error, $\epsilon$ is decreased. When λ is large, then the process is far from optimal value, and it operates like the gradient descent algorithm; and when λ is small, the process is close to optimality and operates like the GN algorithm. In determining λ during each iteration, a constant value (say $v$ = 10) is used to either multiply or divide λ ($\lambda/v \; or \; \lambda \cdot v$), depending on whether the process is closer or farther to the optimal value (Gavin, 2019; Dhkichi et al, 2014).

From $\mathbf{J}^{\mathrm{T}}(\mathbf{J}\delta_{\mathbf{p}} - \epsilon) = 0$, we obtain solution $\delta_{\mathbf{p}}$ of the equation: $\mathbf{J}^{\mathrm{T}}\mathbf{J}\delta_{\mathbf{p}} = \mathbf{J}^{\mathrm{T}}\epsilon$.

For each computation step that leads to a decrease in the error, $\epsilon$, the resultant $\delta_{\mathbf{p}}$ is accepted, and a corresponding decrease in λ is effected as the process is repeated. The resultant $\delta_{\mathbf{p}}$ is rather rejected if $\epsilon$ increased, and λ is consequently increased to continue the process, till the error begins to decrease again. The damping is adjusted at each step of the process until when the stopping criteria are achieved. With these, Loukaris (2005) noted that LM algorithm is adaptive since it controls its damping by itself.

To avoid the problem of slow convergence caused by large λ, the identity matrix **I**in (3.38) was replaced with the diagonal matrix of$\mathbf{J}^{\mathrm{T}}\mathbf{J}$, such that:

$$[\mathbf{J}^{\mathrm{T}}\mathbf{J} + \lambda\,\mathrm{diag}(\mathbf{J}^{\mathrm{T}}\mathbf{J})]\delta_{\mathbf{p}} = \mathbf{J}^{\mathrm{T}}\,[y_i - f(\mathbf{p})] \qquad (3.39)$$

The minimum or optimal value is reached when$\mathbf{J}\delta_{\mathbf{p}} - \epsilon$ is orthogonal to **J**.

The particular parameter to solve for is given by the iterative formula of LM:

$$\mathrm{p}_i^{k+1} = \mathrm{p}_i^{k} - \delta_{\mathbf{p}} \qquad (3.40)$$

From (3.38): $\mathrm{p}_i^{k+1} = \mathrm{p}_i^{k} - (\mathbf{J}^{\mathrm{T}}\mathbf{J} + \lambda\mathbf{I})^{-1}\mathbf{J}^{\mathrm{T}}[y_i - f(\mathbf{p})]$ (3.41)

From (3.39): $\mathrm{p}_i^{k+1} = \mathrm{p}_i^{k} - (\mathbf{J}^{\mathrm{T}}\mathbf{J} + \lambda\mathrm{diag}(\mathbf{J}^{\mathrm{T}}\mathbf{J}))^{-1}\mathbf{J}^{\mathrm{T}}[y_i - f(\mathbf{p})]$ (3.42)

In most cases of parameter computation, the RMSE between calculated and observed results is calculated (Ouadfeul and Aliouane, 2015; Dkhichi et al, 2014).

A number of software have sub-routine codes for the implementation of LM algorithm,such that it is built into software like Matlab, LabVIEW, Mathematica, C/C++, Octave and GNU. Few of the functions for LM are as follows:

- *levmar*, a C/C++ implementation of LM that can be found at http://www.ics.forth.gr/~lourakis/levmar (Lourakis, 2005).
- *lsqcurvefit*, the MATLAB curve-fitting commanddiscussed at Mathwork website: https://de.mathworks.com/help/optim/ug/lsqcurvefit.html?searchHighlight=lsqcurvefit&s_tid=doc_srchtitle (Mathworks, 2019).

  Mostimplementations of LM algorithm are done in MATLAB environment.

### 3.7.4 Running LM Algorithm on the Levmar Platform

The LM algorithmisin Appendix A, while Figure 3.9the process flowchart.

The *Levmar*programme was used for running the non-linear logistic function, in order to obtain optimised values for the curve steepness parameter and the initial function parameter. '*Levmar*' is an ANSI C programme developed to implement the LM non-linear least squares algorithm on C/C++ programming language registered under the GNU General Practice License (GNU-GPL). The software package '*Levmar*' was so chosen for this work because of its versatility and applicability with sub-routines built into many high-level programming languages which include Matlab, Python, and Octave. Incorporated in *Levmar*is an interface file known as MEX-file for interfacing with Matlab, as indicted in the levmar.m file.

Both double and single data precision variants are consisted in *levmar*, denoted by the prefixes 'd' and 's' in the function codes respectively. In compiling *levmar*, only one of the double or the single variant could be specified. Levmar evaluates the Jacobian by numerical finite difference approximations, using Broyden's rank one updates, or by analytic approach and denoted with 'dif' and 'der', respectively, in the codes. Numerical approximation of the Jacobian is relatively slow to converge, but it is adopted particularly when the analytic computation of the Jacobian becomes difficult or expensive. The code: dlevmar_chkjac ( ) or slevmar_chkjac ( ) is used to check the consistency or correctness of the Jacobian with the function.

Depending on the nature of the non-linear problem, the initial values of the parameters are determined by guesses based on background experience and

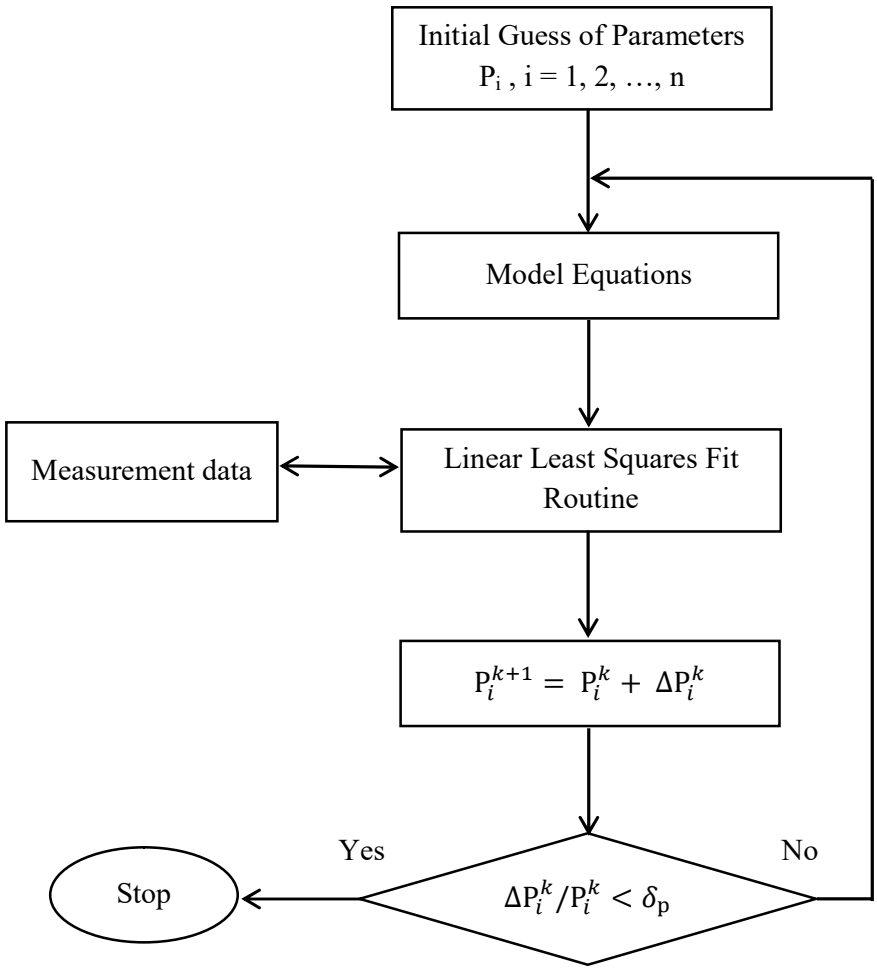



Figure 3.9. Optimisation Flow Chart using Levenberg-Marquardt Algorithm. (Adopted from Duc-Hung et al, 2012).

knowledge of the particular problem, use of the solution obtained from linearising the non-linear function, or use of random values. For this work, solution was obtained by carrying out linear regression through linearisation of the function to obtain a more analytical initial parameter values.

Conditions for terminating the LM algorithm require that at least one of the following should be fulfilled (Loukaris, 2005):

1. The gradient, $\mathbf{J}^{\mathrm{T}}\epsilon$, in (3.38), should drop to less than the threshold, $\varepsilon_1$;
2. The relative change in $\delta_{\mathrm{p}}$ should drop to less than the threshold, $\varepsilon_2$;
3. The error should drop to less than the threshold, $\varepsilon_3$; or
4. The specified maximum number of iterations, $k_{\max}$, is fulfilled.

Parameter values specified in the programme for the iteration are as follows:

1. The initial values of the function parameters;
2. The initial parameter, $\tau$, of the damping factor; and
3. The stopping criteria for convergence, $\varepsilon_1$, $\varepsilon_2$, and $k_{\max}$.

# CHAPTER FOUR
# RESULTS AND DISCUSSIONS

## 4.1 Introduction

Results of all activities carried out during this study are stated and discussed in this chapter. These include results of speech acquisition, conversion and transmission on selected wireless mobile networks. Results and analyses are arranged according to the objectives of these research efforts.

Results of works done on research objective one: study and analysis of the psychoacoustic parameters of both original and received speechesare in Section 4.3. Results on research objective two: perceptual quality tests carried out using subjective listening-only test approach are in Section 4.4. Results of works on research objective three: perceptual quality tests carried out using intrusive objective test approach are in Section 4.5. Results on research objective four: research efforts that led todeveloping an improved mapping function for mapping the raw quality score of PESQ model to the ideal subjective listening-only MOS scale are in Section 4.6. Also included are results of correlational analysis and calculations of correlation errors for these quality test approaches and works.

## 4.2 Speech Recording, Conversion and Transmission

Original speeches were recorded using the CUBASE software in the 'amr' format and were converted to the 'wav' format for processing using the 'Any Audio Converter (AAC)' software. A sample of the recorded and converted speech before it was received is shown in Figure 4.1. The speech in Figure 4.1 is (OF1S4): "The Senator embezzled the country's money … and proved innocence after the accusation." It was read by a young lady of about 22 years old.

The speech after conversion is in 'wav' format, and the waveform as shown in Figure 4.1 is the temporal structure of a sample original speech signal. Itconsists of the two parts with each part occupying approximately 2.5 seconds duration. The speech sample has 1 second silence at the beginning, the two speech parts are separated by about 1.3 seconds, and the second part of the speech ended with a silence

duration of about 1.3 seconds. This is temporal speech structure is in keeping with the ITU-T guidelines for recorded speech for such research study as this.

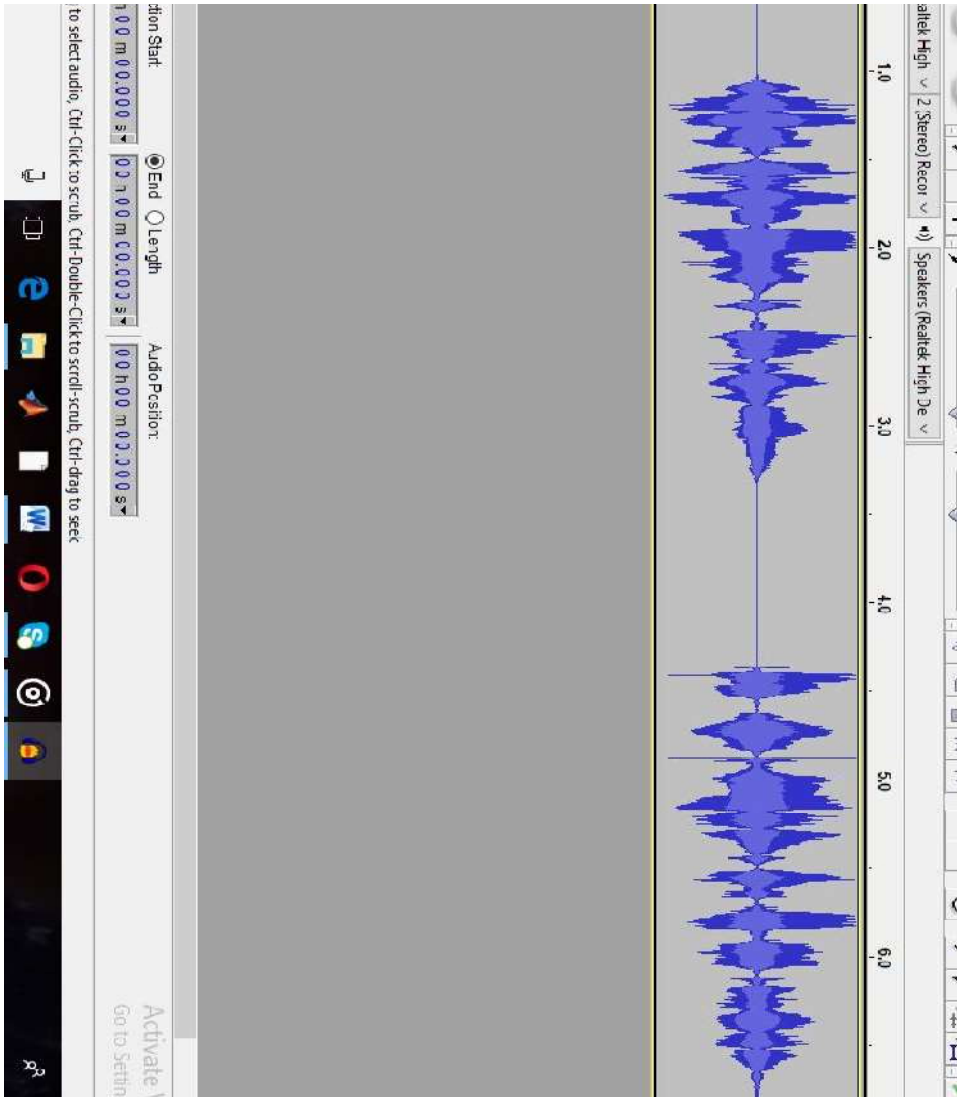


Figure 4.1. Plot of the Temporal Structure of aSample Recorded and Converted Speech (OF1S4)beforeTransmission.

The 64original speech files in 'wav' speech format and the reading of the Sound Pressure Level (SPL) meter (in dBA) while each speech was being recorded are listed in Table B.1 of the Appendix. The 64 speeches consist of 32 male speeches and 32 female speeches.

The 64 original speeches were transmitted over three mobile wireless telephone networks in Nigeria made up of two intra-networks (Networks A, and B) and one inter-network (Network C).The received speeches were recorded by the receiving phone handset using the Call Recorder application, are listed in Tables B.2

to B.4 of the Appendix. Transmission over three networks, gave us total of 192 received speeches.

## 4.3 Psychoacoustic Parameters of Speech and Speech Quality

The results of evaluating key psychoacoustic parameter of speechthat bothers on quality of received or degraded speech began with signal analysis of sample original and received speech signals. This was followed with obtaining loudness parameters or features of these speech signals, as discussed in this section. The phenomena of loudness of both the original and received speeches were considered. With a couple of loudness programming, an innovative intrusive approach of comparatively deducing quality of degraded speech from the loudness scores was developed.

### 4.3.1 Waveform Analysis Plots of Original and Received Speeches

Shown in Figure 4.2 is the waveforms (in dB) plot of original speech sample OM1S1.wav and its correspondingreceived speeches transmitted overthe three different wireless networks – two intra-networks (Networks A and B) and one inter-network (Network C), that is, AM1S1.wav,BM1S1.wav, and CM1S1.wav. This waveform plots were carried out on Audacity software (Audacity 2.1.2., 2015).

### 4.3.2 Frequency Analysis Plots of Original and Received Speeches

Figures 4.3 to 4.6 are the plots of the frequency analysis (spectrum)of the original sample speech and the corresponding received speeches from the three networks, carried out on Audacity. These also showed the robustness of the spectrum of the original speech compared to the frequency spectrum plots of its corresponding received speeches shown in Figures 4.4, 4.5 and 4.6.

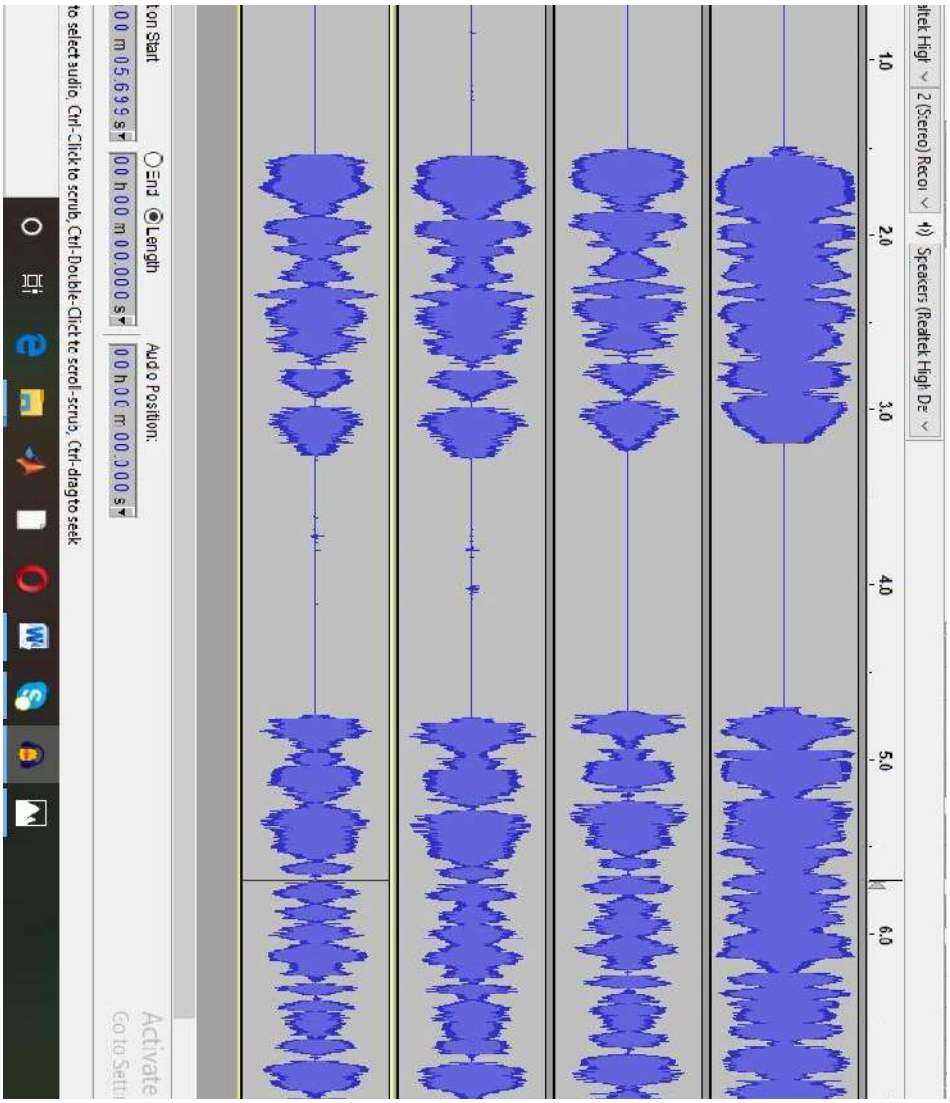



Figure 4.2. Plot of the temporal structurefor (a) original speech signal – OM1Sp.wav (b) Speech over NetworkA (c) Speech over NetworkB (d) Speech over Network C.

### 4.3.3 Spectral Analysis onSample Speeches

The Hanning window function given by:

$$w(n) = 0.5\left(1 - cos\left(\frac{2\pi n}{N-1}\right)\right) = hav\left(\frac{2\pi n}{N-1}\right) \quad (4.1)$$

was applied in stratifying the speech signals to obtain the frequency and spectral values for the spectral analysis plots for the speeches. These values were read off the log frequency plots of the original and received speeches and tabulated on Table 4.1. These values areused for the plots of spectral analysis, the intensities or spectral amplitude against the frequency as shown in Figure 4.7.

These speech signals were sampled at the rate of 44100 Hz and digitized (PCM) on a 16-bit format. The fast Fourier transform (FFT) analyser of the Audacity

software applies the Hanning window function to the sampled signal by multiplying the digitized signal by the Hanning window function and was displayed in Figure 4.7.

The spectral plot of the original speech signal in Figure 4.7 is seen to be much robust than the plots of the received speech signals. This further shows the attenuation and loss of quality suffered by the received speeches during transmission.

### 4.3.4 Programming of Loudness Estimation

Computing loudness parameters of time-varying sounds is very complex, therefore in order to reducestress, time and resources, most models have one form of computer programme or the other developed for them. The software produced by GENESIS was used on the platform of MATLAB to implement the algorithms of the Zwicker and Fastl model. The result of the computationof the instantaneous loudness parametersfor the sample speech (OM1S1.wav)and its corresponding received speech used for this analysis are stated in Table 4.2.

Comparing the values of the instantaneous loudness for the original and the received speeches and also comparing the loudness level for the original and thereceived speeches in Table 4.2 it could be seen that the maximum instantaneous loudness of the received speeches are 42.55, 37.08 and 35.64% respectively of that of the original speech. It could also be seen that the maximum instantaneous loudness level of the received speeches are 87.06, 84.98 and 84.37% respectively of that of the original speech. These are stated in Tables4.3 and 4.4 respectively.

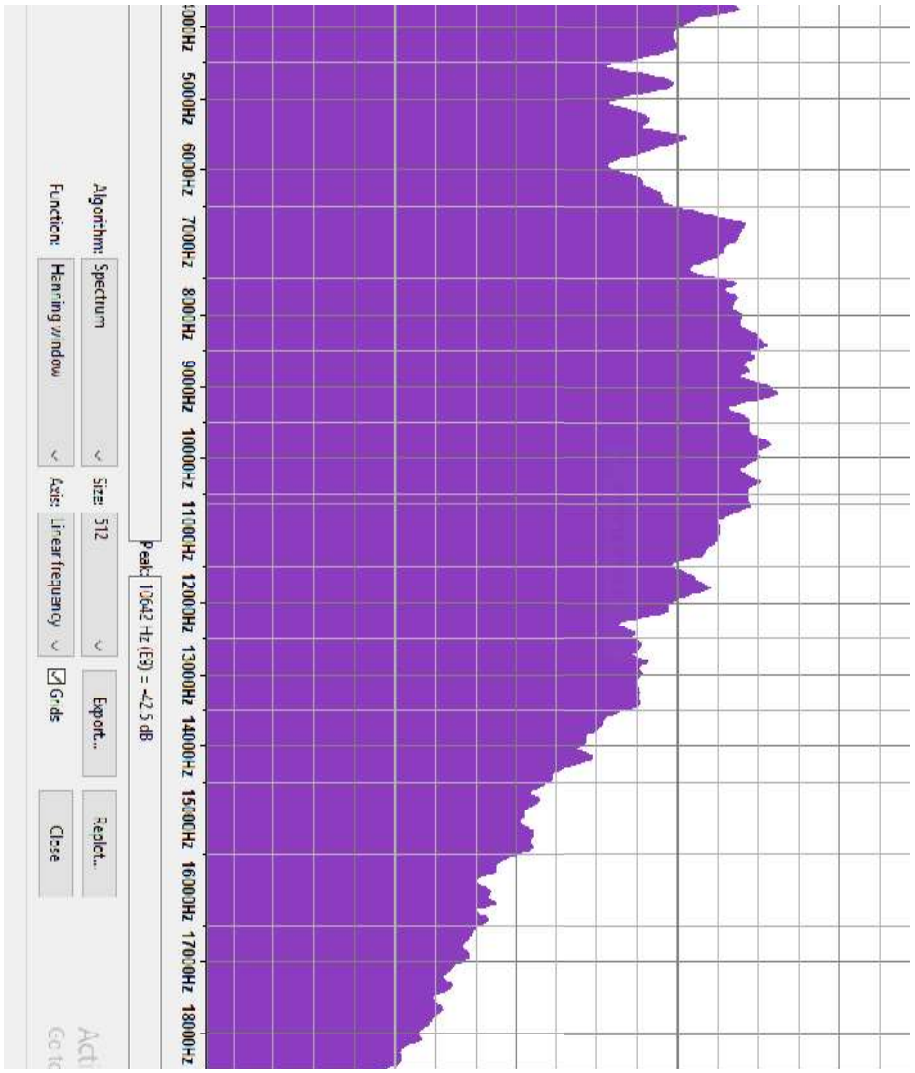



Figure 4.3. Spectral Plot of Original Speech OM1S.wav.

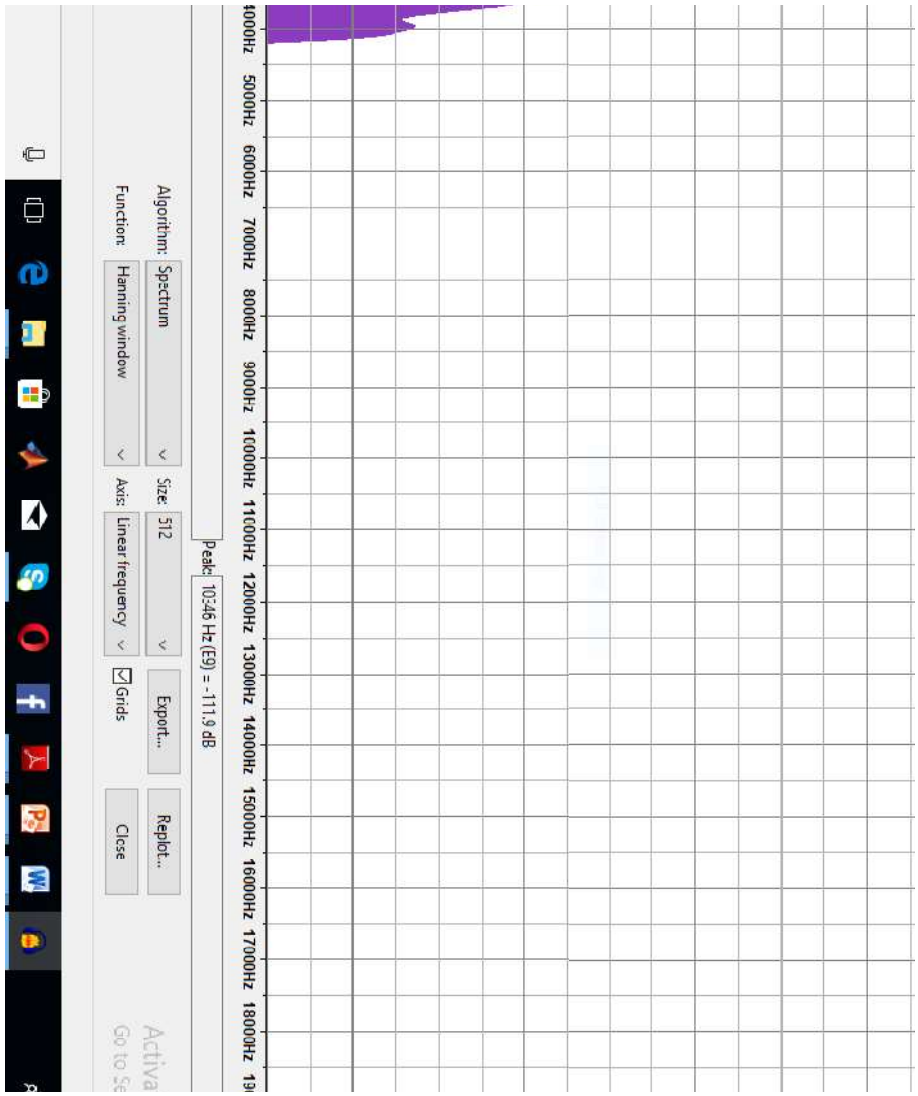



Figure 4.4. Spectral Plot ofReceived Speech from Network A (AM1S1.wav).

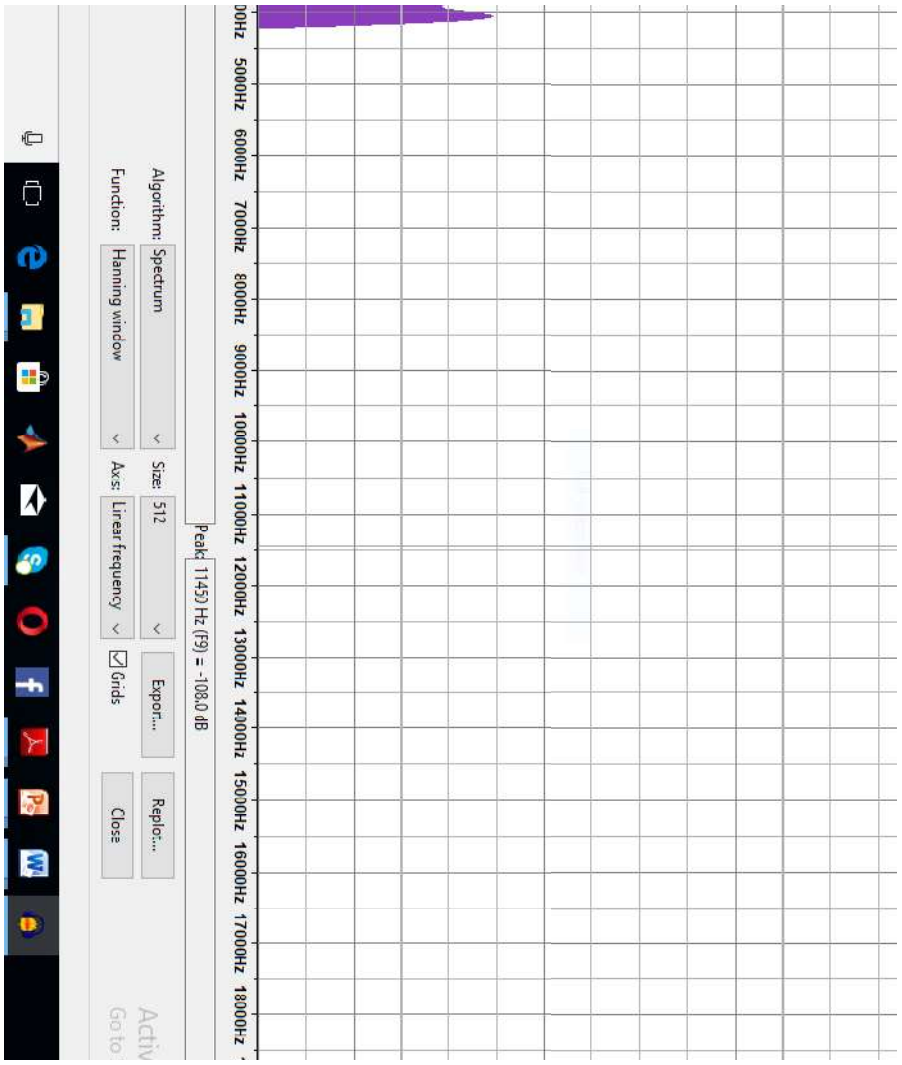



Figure 4.5. Spectral Plot ofReceived Speech fromNetwork B (BM1S1.wav).

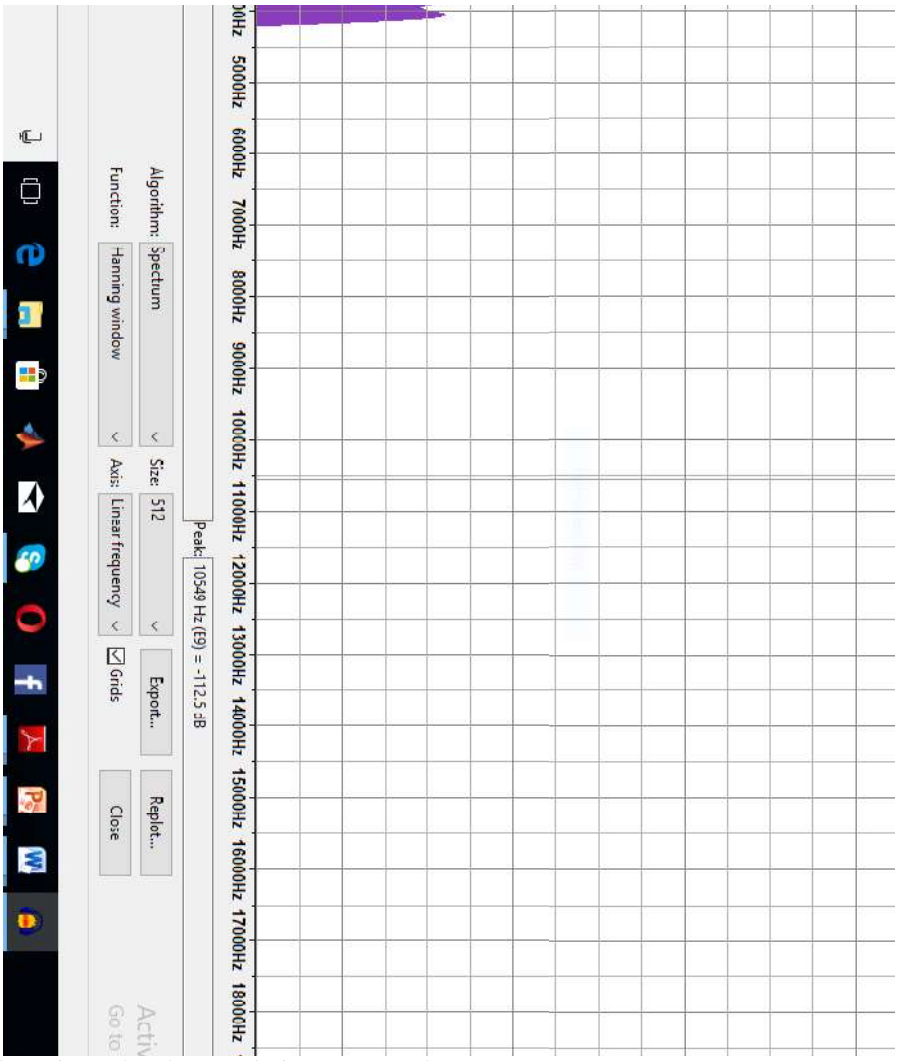



Figure 4.6. Spectral Plot ofReceived Speech fromNetwork C (CM1S1.wav).

Table 4.1. Table of the Spectral Values versus Frequency of Original and ReceivedSpeeches.

| Frequency (Hz) | Spectra Level (dB) | | | |
|---|---|---|---|---|
| | OM1S1 | AM1S1 | BM1S1 | CM1S1 |
| 43.0664 | -55.3385 | -80.2051 | -78.4764 | -79.6087 |
| 86.1328 | -40.6508 | -72.6653 | -72.2935 | -72.5652 |
| 129.1992 | -35.3355 | -69.8775 | -70.4776 | -70.1223 |
| 172.2656 | -34.9163 | -69.5041 | -69.5272 | -68.6655 |
| 215.3320 | -30.7254 | -66.1885 | -67.6960 | -67.5588 |
| 301.4648 | -26.0986 | -55.8477 | -61.3761 | -61.6812 |
| 430.6641 | -31.1379 | -52.9308 | -52.0475 | -49.4478 |
| 516.7969 | -32.6376 | -48.6912 | -49.3775 | -46.2927 |
| 602.9297 | -35.1511 | -50.7529 | -51.9424 | -48.2852 |
| 732.1289 | -35.8761 | -53.1410 | -50.9537 | -50.2957 |
| 818.2617 | -39.5773 | -54.1640 | -51.8773 | -54.9020 |

| | | | | |
|---|---|---|---|---|
| 904.3945 | -43.7969 | -51.3297 | -48.9996 | -51.8925 |
| 1033.5938 | -43.2288 | -48.8864 | -49.6921 | -48.9766 |
| 1119.7266 | -43.9858 | -48.0442 | -48.8646 | -49.3255 |
| 1205.8594 | -43.1637 | -43.7477 | -45.7731 | -51.6784 |
| 1335.0586 | -40.5850 | -39.4985 | -41.8301 | -52.2926 |
| 1421.1914 | -37.7834 | -40.8600 | -41.4215 | -52.2299 |
| 1507.3242 | -41.3901 | -42.2119 | -43.9049 | -54.1514 |
| 1636.5234 | -40.2303 | -45.0955 | -48.4650 | -54.1394 |
| 1722.6563 | -41.4034 | -44.9353 | -45.7151 | -56.1189 |
| 1808.7891 | -43.1211 | -44.2251 | -43.4640 | -59.8371 |
| 1937.9883 | -41.7084 | -50.3428 | -45.0490 | -54.9690 |
| 2024.1211 | -45.9009 | -53.3532 | -45.3526 | -51.8424 |
| 2153.3203 | -46.2074 | -55.1391 | -47.0198 | -49.9878 |
| 2239.4531 | -45.1717 | -58.6082 | -44.1911 | -48.4161 |
| 2325.5859 | -46.6965 | -60.1810 | -45.7958 | -50.3560 |
| 2411.7188 | -47.3157 | -60.7946 | -46.7988 | -51.8370 |
| 2540.9180 | -49.6568 | -62.5041 | -49.6358 | -53.7152 |
| 2627.0508 | -49.5794 | -61.2902 | -55.2876 | -55.1217 |
| 2713.1836 | -47.9161 | -58.3059 | -56.7108 | -54.1423 |
| 2842.3828 | -53.1400 | -55.6363 | -52.3514 | -50.9596 |
| 2928.5156 | -55.1181 | -52.4669 | -52.7525 | -48.6873 |
| 3014.6484 | -55.0208 | -49.5796 | -53.2773 | -47.6358 |
| 3100.7813 | -53.0007 | -49.3162 | -53.5450 | -48.7805 |
| 3229.9805 | -49.9421 | -49.7172 | -45.7324 | -45.2961 |
| 3316.1133 | -50.2388 | -56.2672 | -52.4298 | -50.7579 |
| 3402.2461 | -49.4628 | -63.1617 | -62.3250 | -62.5208 |
| 3531.4453 | -43.6770 | -76.4379 | -68.3834 | -71.1503 |
| 3617.5781 | -48.9158 | -73.8147 | -72.9104 | -75.0457 |
| 3703.7109 | -48.0975 | -75.4344 | -74.8157 | -76.1662 |
| 3832.9102 | -49.2664 | -85.9536 | -78.1619 | -80.5395 |
| 3919.0430 | -50.3070 | -82.1564 | -80.2007 | -80.7120 |
| 4005.1758 | -51.6602 | -82.1839 | -81.6645 | -82.4104 |
| 4134.3750 | -51.7549 | -85.9979 | -81.9793 | -81.5411 |
| 4521.9727 | -56.6758 | | | |
| 5038.7695 | -58.2574 | | | |
| 5512.5000 | -50.8984 | | | |
| 6029.2969 | -56.1649 | | | |
| 6503.0273 | -52.0434 | | | |
| 7019.8242 | -48.2693 | | | |
| 7536.6211 | -48.3397 | | | |
| 8010.3516 | -47.1280 | | | |
| 8527.1484 | -48.2414 | | | |
| 9000.8789 | -43.8982 | | | |
| 9517.6758 | -45.7672 | | | |
| 10034.4727 | -45.2340 | | | |

| | |
|---|---|
| 11025.0000 | -47.9163 |
| 12015.5273 | -53.1094 |
| 13006.0547 | -54.1308 |
| 14039.6484 | -58.8544 |
| 15030.1758 | -63.5203 |
| 16020.7031 | -65.7298 |
| 17011.2305 | -67.2641 |
| 18001.7578 | -71.7818 |
| 19035.3516 | -76.4991 |
| 20025.8789 | -79.8268 |
| 21016.4063 | -82.7812 |
| 21533.2031 | -84.7335 |
| 21619.3359 | -84.0777 |
| 21705.4688 | -87.0263 |
| 21834.6680 | -86.3106 |
| 21920.8008 | -87.2458 |
| 21963.8672 | -88.1221 |
| 22006.9336 | -85.9968 |

**Notes:** **Format = 32 bit float**
**Rate = 41000 Hz**
**Spectrum Size = 1024**
**This is an abridged version of data generated from Spectra Plot software.**

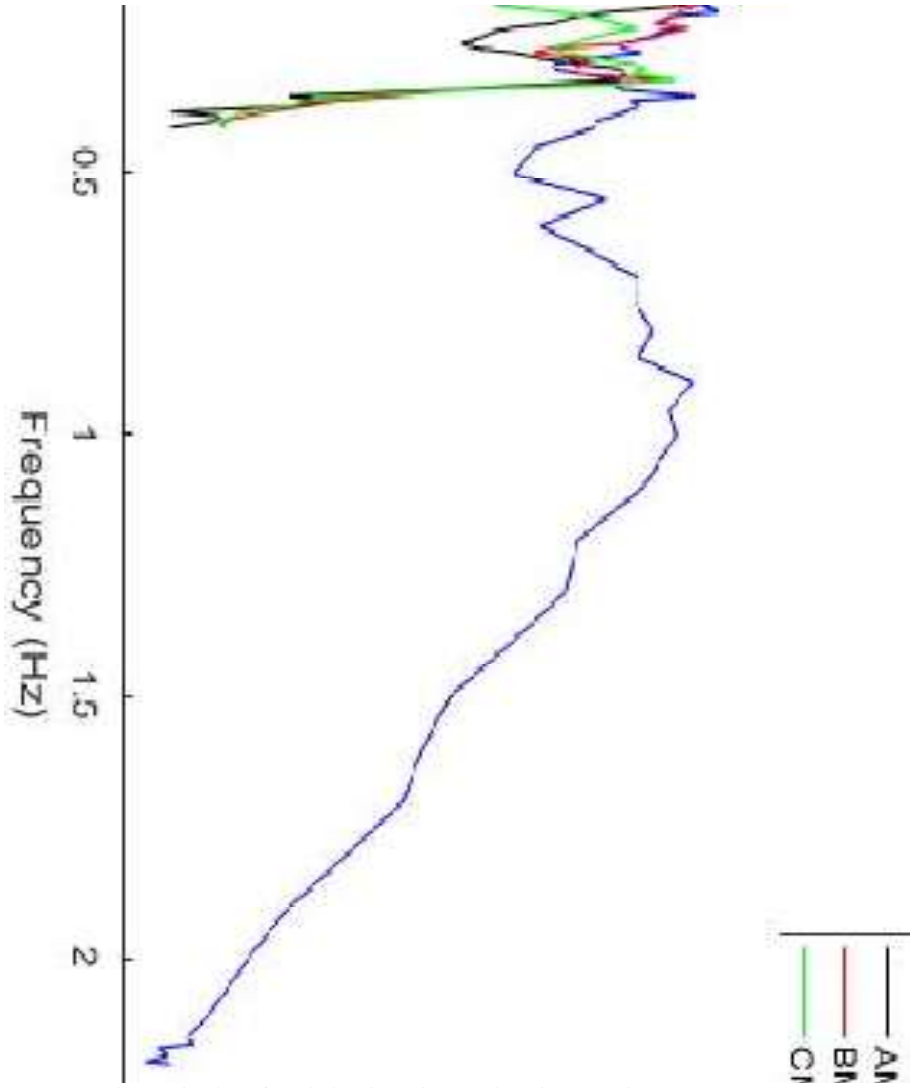



Figure 4.7. Plot of the Spectral Analysis of Original and Received Speeches.

Table 4.2. Instantaneous Loudness and Loudness Level for Original and Degraded Speeches.

| **Speech samples** | | **OrgM1Sp1** | **EEM1Sp1** | **GGM1Sp1** | **EMM1Sp1** |
|---|---|---|---|---|---|
| Instantaneous Loudness | Nx | 32.7322 | 15.6137 | 13.7664 | 12.8566 |
| | Nt | 45.4268 | 19.2678 | 16.6347 | 15.7504 |

| | | | | | |
|---|---|---|---|---|---|
| (sone) | Nmax | 46.1854 | 19.6506 | 17.1257 | 16.4626 |
| Instantaneous Loudness Level (phon) | Lx | 90.3264 | 79.6474 | 77.8308 | 76.8443 |
| | Lt | 95.0547 | 82.6812 | 80.5612 | 79.7731 |
| | Lmax | 95.2936 | 82.9650 | 80.9808 | 80.4112 |

Table 4.3. Comparison of Maximum Instantaneous Loudness for Original and Received Speeches

| **Speech samples** | **Maximum instantaneous loudness (sone)** | |
|---|---|---|
| | **Nmax** | **%Ref** |

| | | | |
|---|---|---|---|
| Original Speech | OM1S1 | 46.1854 | 100 |
| Received Speeches | AM1S1 | 19.6506 | 42.55 |
| | BM1S1 | 17.1257 | 37.08 |
| | CM1S1 | 16.4626 | 35.64 |

Table 4.4. Comparison of Maximum Instantaneous Loudness Levelsfor Original and Received Speeches.

| **Speech samples** | **Maximum instantaneous loudness level (phon)** | |
|---|---|---|
| | **Lmax** | **%Ref** |

| | | | |
|---|---|---|---|
| Original Speech | OM1S1 | 95.2936 | 100 |
| Received Speeches | AM1S1 | 82.9650 | 87.06 |
| | BM1S1 | 80.9808 | 84.98 |
| | CM1S1 | 80.4112 | 84.38 |

## 4.4 Results of Subjective Listening-only Tests Scores

Using ACR subjective technique with listening-only approach, results of average opinion ratings of subjects that listened to the received speeches are listed in Tables C.1, C.2 and C.3 in the Appendix for the three networks.

To analyse the spread of the subjective quality score (MOS), the variances and standard deviations were obtained for the three speech transmission networks and

shown in Tables 4.5 and 4.6. As shown on the two tables, the inter-network speech transmissions (Network C) has wider spread than the two intra-network transmissions, and as well more deviated from the central tendency than the two intra-network speech transmissions.

## 4.5 Results of Intrusive Objective Quality Tests

The Perceptual Speech Evaluation Quality (PESQ) algorithm shown in Figure 3.6 was adopted for the evaluation of the quality of each received speech over the three chosen networks.

Objective quality test scores were obtained inrunning each of the 64 original speechesandits corresponding received speechesthrough the perceptual computation of the PESQ algorithm in accordance with the provisions of ITU-T Rec. P.862. Raw PESQ quality scores within the value range of –0.5 to 4.5 were obtained for the tests. This quality score values were recorded and tabulated in Tables D1 to D3 in the Appendix for the three networks respectively.

### 4.5.1 Results of Mapped Speech Quality Scores

The subjective MOS quality score range provides the true quality score in any quality test for degraded or distorted speeches. The obtained raw PESQ scores were run through the internationally standardized ITU-T Rec. P.862.1 mapping function so as to translate them to the MOS scale to denote the actual or true quality value for each received speech, as stated below for emphasis:

5.0 stands for Excellent Speech Quality

4.0 stands for Good Speech Quality

3.0 stands for Fair Speech Quality

2.0 stands for Poor Speech Quality

1.0 stands for Bad Speech Quality

Table 4.5. Variance of Subjective Quality Scores for the Received Speeches.

| Network | A | B | C | REMARK |
|---|---|---|---|---|
| Male | 0.143086 | 0.12749 | 0.418506 | Network C (an inter-network) transmission has wider spread than the two intra-network transmissions |
| Female | 0.134287 | 0.249961 | 0.302500 | |
| Overall | 0.142029 | 0.199685 | 0.374236 | |

Table 4.6. Standard Deviation of Subjective Quality Scores for the Received

Speeches

| Network | A | B | C | REMARK |
|---|---|---|---|---|
| Male | 0.38432 | 0.357058 | 0.646929 | Network C (an inter-network) transmission more deviated spread than the two intra-network transmissions |
| Female | 0.378267 | 0.499961 | 0.550000 | |
| Overall | 0.379846 | 0.446861 | 0.611748 | |

The mapped objective intrusive quality score for each received speech is given on Tables E.1 to E.3 in the Appendix for the three networks as PESQ MOS-LQO score using ITU-T Rec. P.862.1 mapping function given by:

$$y(x) = 1 + \frac{4.999 - 0.999}{1 + e^{(-1.4945x + 4.6607)}} \quad (4.2)$$

where $x$ is the raw PESQ score obtained for each received speech and $y(x)$ is the ideal quality score.

### 4.5.2 Scatter/Regression Plots

In Figures 4.8 to 4.10, we have the scatter plots andthe results of the linear regression analysis of the mapped objective quality scores (MOS-LQO) of the speech signalsreceived from the three Networks A, B, and C, respectively.

### 4.5.3 Results of Statistical Analyses

#### *1. Correlational Analysis*

The closeness of fit between the objective intrusive test scores MOS-LQO and the subjective MOS-LQS obtained by calculating the correlation coefficient using the Pearson's formula. The Pearson's correlation coefficient is the main measure of the performance of objective models and the most common metric for evaluating the performance of objective speech quality estimation methods between subjective and objective test values as noted by (Kim and Tarraf, 2004). The Pearson's formula is given by:

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{4.3}$$

where, $x_i$ is the individual Subjective MOS value, $\bar{x}$ is the mean over all the subjective MOS values for a particular transmission, $y_i$ is the individual PESQ MOS-LQO value, and $\bar{y}$ is the mean over all the PESQ MOS-LQO values for a particular transmission.

The results of obtaining the correlation coefficientsbetween the PESQ MOS-LQO and the Subjective MOS for speeches transmission over the three networks are given in Table 4.7.

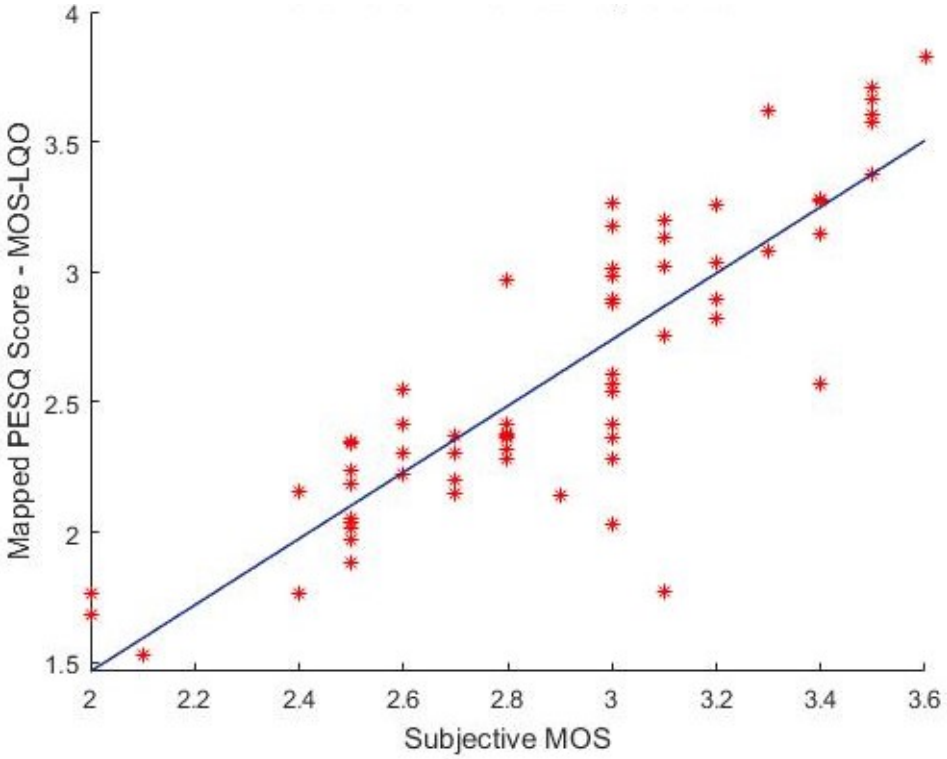



Figure 4.8. Scatter/Regression Plot of Network AReceived Speeches.

***The Regression function, R,***was obtained as: $y(x) = 1.2751x - 1.0850$

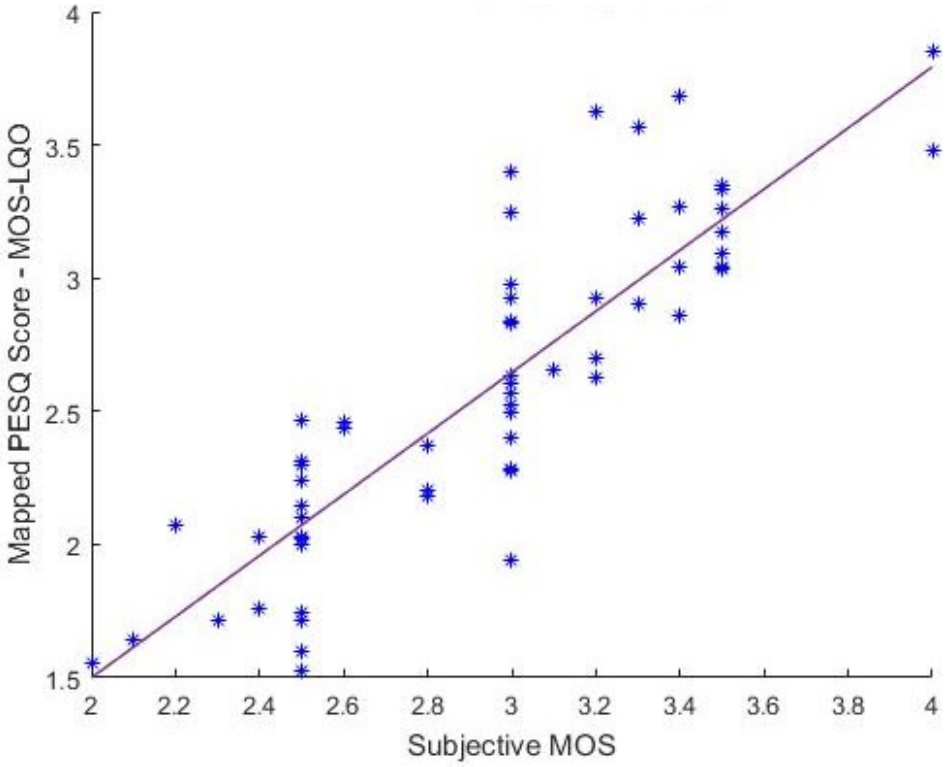



Figure 4.9. Scatter/Regression Plot of Network BReceived Speeches.

***The Regression function, R,***was obtained as: $y(x) = 1.14823x - 0.8001$

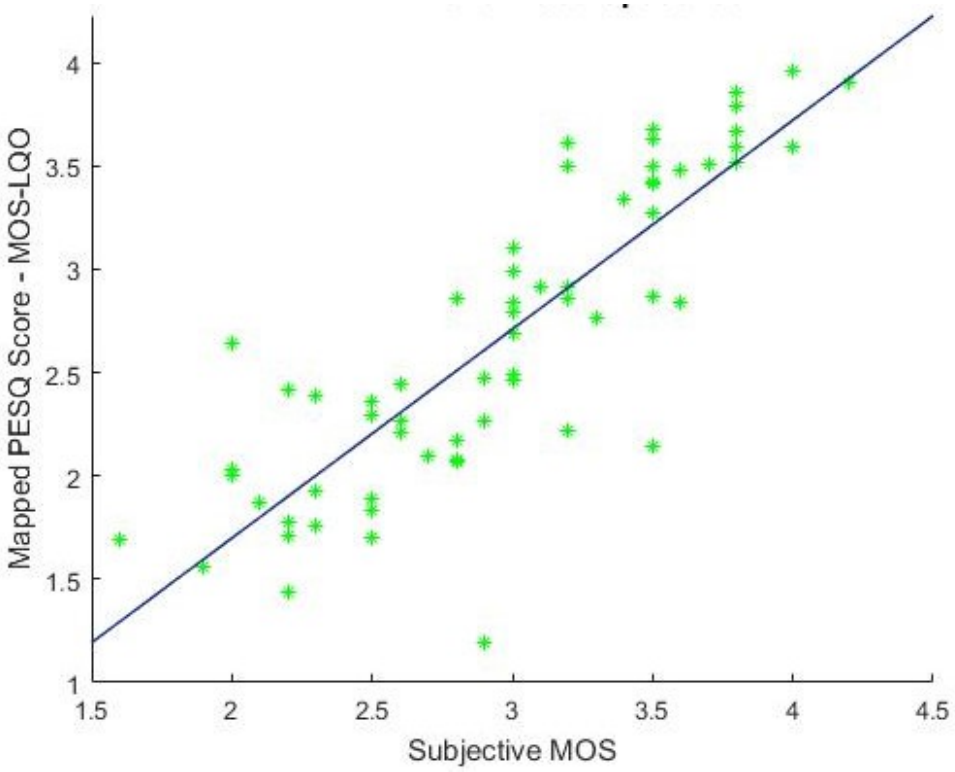



Figure 4.10. Scatter/Regression Plot of Network CReceived Speeches.

***The Regression function, R,***was obtained as: $y(x) = 1.0075x - 0.3035$

***2. Root mean square error (RMSE)***

Furthermore, comparing the results of correlation and root mean square error with that of the ITU-T standard for objective speech quality assessment provides a means to determine the efficiency of the adopted method and/or designed algorithm.

MOS measurement accuracy is assessed using the root-mean-square MOS error (*RMSE*) given by (Wang et al, 2008):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \tag{4.4}$$

where *N* is the number of MOS labeled utterances used in the evaluation.

The results of the root mean square calculation for the three networks are given on Table 4.8.

### *3. Prediction error, $E_P$*

Provides the average standard error of the objective estimator of the subjective score, that is, the average evaluation error and given by (Cotanis, 2009):

$$E_P = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N-1}} \tag{4.6}$$

The prediction error obtained for Networks A, B, and C, respectively, are given on Table 4.9.

## 4.6 Results of Optimising Logistic Function Parameters

Efforts at obtaining optimal values of parameters, *b* and *c*, steepness of the curve and the constant of integration that represents the initial point of the independent variable of the logistic function:

$$y(x) = 1 + \frac{4}{1 + e^{-(bx+c)}} \tag{4.9}$$

took into consideration the boundary and range conditions stated below:

$x$: Scale of raw PESQ algorithm output: -0.5 to 4.5

$y(x)$: Scale of ideal Subjective MOS scores: 1.0 to 5.0

**Extreme Cases:**

Case 1:

Substituting $(x, y) = (-0.5, 1.0)$, into the function results in error, that is:

$$e^{0.5b-c} = -1, \text{and } 0.5b - c = \ln(-1) = error \tag{4.10}$$

Table 4.7. Correlation Coefficients for the Subjective vs. PESQ MOS-LQO.

| Network | A | B | C |
|---|---|---|---|
| Correlation Coefficient, R | 0.854 | 0.871 | 0.848 |

Table 4.8.RMSE for the Subjective vs. PESQ MOS-LQO.

| Network | A | B | C |
|---|---|---|---|
| Root Mean Square Error (RMSE) | 0.4230 | 0.4687 | 0.4787 |

Table 4.9.Prediction Errorsfor the Subjective vs. PESQ MOS-LQO.

| Network | A | B | C |
|---|---|---|---|
| Prediction error, $E_P$ | 0.4264 | 0.4724 | 0.4825 |

Case 2:

Substituting $(x, y) = (4.5, 5.0)$ into the function also results in error:

$$e^{-(bx+c)} = 0$$

This is because, $y$ is never equal to 5.0 at $x = 4.5$.

$y = 5.0$, only at $x = +\infty$ according to the limit:

$$\lim_{x \to +\infty} y(x) = 5.0$$

These cases establish the asymptotic nature of the logistic function both at the starting part and at the peak.

Results of Non-linear least squares problem:

Result of the non-linear least squares regression of the logistic function were obtained from the following and used as initial values for the optimisation process:

Measured data on Tables C.1 to C.3 in the Appendix for subjective scores and Tables D.1 to D.2in the Appendix for objective PESQ scores, were transformed by linearisation as discussed in Sub-section 3.7.2.The transformed data were plotted on MATLAB. The result of the least squares regression parameters is given on Table 4.10.

The Levenberg-Marquardt optimisation software, levmar, was run on and compiled with Dev C++ compiler. Initial parameter, maximum iteration and error thresholds stopping criteria specified for the process, and the result of the process are specified on Table 4.10.

Substituting optimised parameter values into the logistic function we have:

$$y(x) = 1 + \frac{4}{1 + e^{-2.2106x \quad .5781}} \tag{4.11}$$

Figure 4.11 shows the plot of this new function.

### 4.6.1 Comparison oflogistic mapping functions

Correlation coefficient of the derived function was calculated using existing data of received speeches on tables, resulted in 0.849. The coverage of the MOS scale using this function was also computed and was compared with those of the two knownstandard mapping functions, namely: the ITU-T Rec P.862.1 and the United States patented Morfitt III and Cotanis logistic mapping function. The results are shown on Table 4.11and the plot of the three functions is shown in Figure 4.12.

Table 4.10: Results of Regression and Optimisation Processes.

| **Category** | **Parameter** | **Value** |
|---|---|---|
| Initial parameter values (obtained from linearisation of data for nonlinear regression) | Steepness parameter, $b$ | 1.3 |
| | Parameter of integration constant, $c$ | 0.1225 |
| Values specified for the optimisation process | Error thresholds, $\varepsilon_1$ | $10^{-8}$ |
| | Error thresholds, $\varepsilon_2$ | $10^{-8}$ |
| | Maximumnumber of iteration, $k_{max}$ | 120 |
| | Initial value of the damping factor, $\tau$ | $10^{-2}$ |
| Results of the optimisation process: | Steepness parameter, $b$ | 2.2106 |
| | Parameter of integration constant, $c$ | -5.5781 |
| | Stopping number of iteration | 35 |

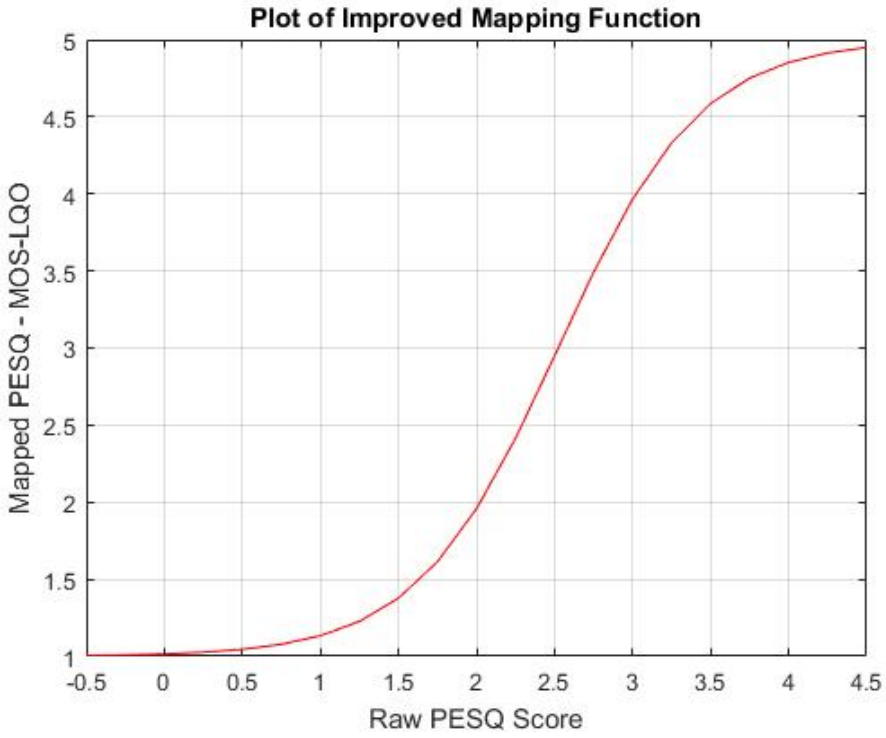



Figure 4.11. Plot of Obtained Logistic Function.

Table 4.12. Comparing Obtained Mapping Function with two Prominent Functions.

| S/N | Raw PESQ Score | Subjective MOS Score | ITU-T Rec. P.862.1 mapped PESQ MOS Score | U. S. Patented logistic function mapped PESQ MOS Score | Obtained logistic function mapped PESQ MOS |
|---|---|---|---|---|---|
| 1. | -0.5 | 1 | 1.077321721 | 1.011137984 | 1.00499980 |
| 2. | 4.5 | 5 | 4.548638319 | 4.757634956 | 4.95000751 |
| Difference between highest & lowest scores | | 4 | 3.471316598 | 3.746496972 | 3.94500771 |
| %age of MOS Score | | 100% | 86.8% of MOS | 93.7% of MOS | 98.6% of MOS |

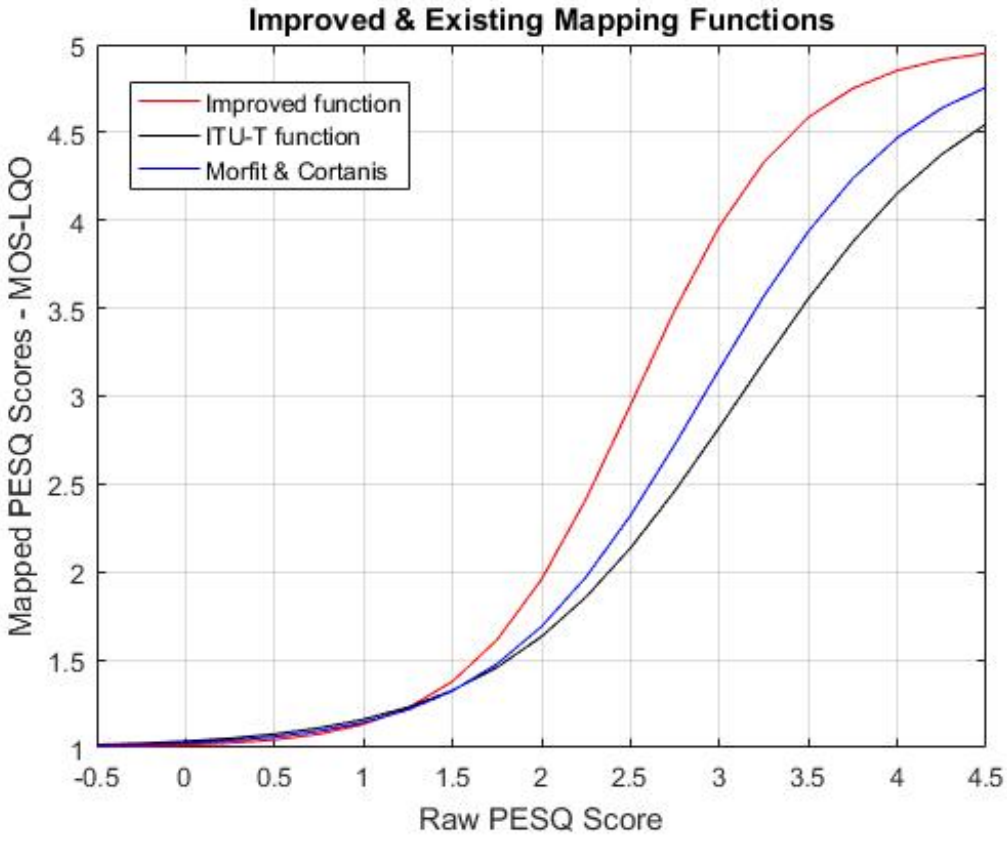



Figure 4.12. Comparison of Obtained Logistic Functions with Existing Ones.

### 4.6.2 Hypothesis testing of the logistic (mapping) functions

Putting obtained logistic mapping function side-by-side the other two functions, we evaluated the variability in their mapped data by testing the following hypotheses using analysis of variance (ANOVA):

Null hypothesis: $H_0$: $\mu_1 = \mu_2 = \mu_3$

Alternate hypothesis: $H_1$: $\mu_l \neq \mu_m$

The null hypothesis is true only if all sample means of the mapped data distributions are equal or do not have any significant differences, while the alternate hypothesis on the other hand, is only true when at least one of the sample means is different from the remaining sample means.

The received speeches from Network A and the corresponding raw PESQ scores inclusive of the mapped data for the three logistic (mapping) functions under test were used as the sample data for ANOVA testas given on Table F.1 in the Appendix.

The mean of the data distribution of the three logistic functions were obtained as: $\mu_1 = 2.614813, \mu_2 = 2.817984$ , $\mu_3 = 3.751563$ and $\mu_G = 3.061453$

The sum-of-squares for between group variability was obtained as:

$$SS_{between} = N1(\mu_1 - \mu_G)^2 + N2(\mu_2 - \mu_G)^2 + N3(\mu_3 - \mu_G)^2$$
$$= 47.04101$$

The degree of freedom between group variability was obtained as:

$$df_{between} = j - 1 = 3 - 1 = 2$$

The mean square for between group variability was obtained as:

$$MS_{between} = \frac{SS_{between}}{df_{between}} = 23.520505$$

The sum-of-square for within group variability was obtained as:

$$SS_{with} = \sum (x_{i1} - \mu_1)^2 + \sum (x_{i2} - \mu_2)^2 + \sum (x_{i3} - \mu_3)^2$$
$$= 73.35076$$

The degree of freedom for within group variability was obtained as:

$$df_{within} = (n_1 - 1) + (n_2 - 1) + (n_3 - 1) = 63 \times 3 = 189$$

The mean square within group variability was obtained as:

$$MS_{with} = \frac{SS_{with}}{df_{with}} = \frac{73.35076}{189} = 0.3881$$

The F-statistic was obtained as:

$$\mathrm{F} - statistic = \frac{Between\ Group\ Variability}{Within\ Group\ Variability} = 60.604239$$

Here we have F-statistic ratio of 60.604 with degree of freedom of (2, 189). The density plot of this degree of freedom is shown in Figure 4.13. Also, for most hypothesis tests, value of alpha, the significance level, the standard is usually taken as the $\alpha = 0.05$.

Summary of the ANOVA test results are given in Tables 4.11 and 4.12 and a density plot of the results was carried out and shown in Figure 4.13.

Table 4.12. Summary of Results of Hypothesis tests.

| *Groups* | *Count* | *Sum* | *Average* | *Variance* |
|---|---|---|---|---|
| ITU-T P.862.1 mapping function | 64 | 167.348 | 2.614813 | 20.28194 |
| Morfitt III & Cotanis mapping function | 64 | 180.351 | 2.817984 | 26.76211 |
| Obtained mapping function | 64 | 240.1 | 3.751563 | 26.30671 |

Table 4.13. Results of ANOVA

| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F-crit* |
|---|---|---|---|---|---|---|
| Between groups | 47.04101 | 2 | 23.52056 | 0.6042 | | |
| Within groups | 73.35076 | 189 | 0.3881 | | | |
| Total | 120.39177 | 187 | | | | 19.49 |

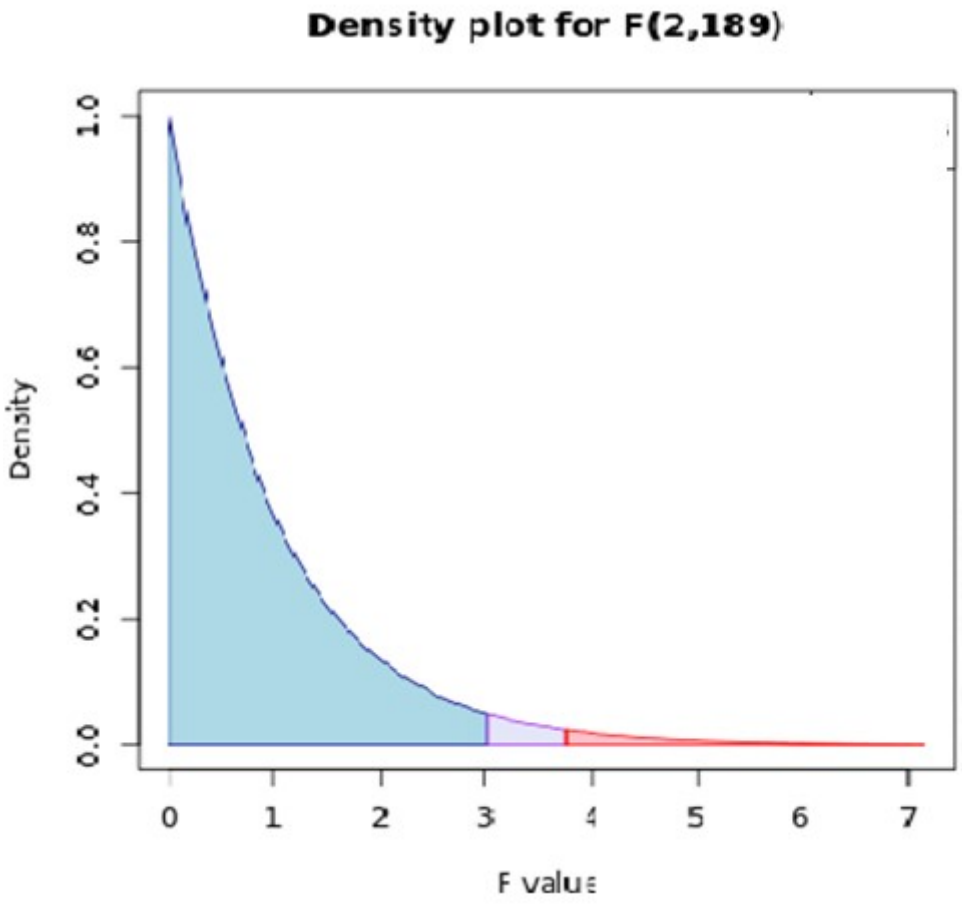



Figure 4.13. Density Plot of the ANOVA Test.

## 4.7 Discussion of Results

### 4.7.1 Temporal Structures of Original and Received Speech Signals

The plots of the temporal structures (waveform) of the sample original speech signal (OM1S1) and its corresponding received speeches: AM1S1, BM1S1 and CM1S1, from networks A, B, and C respectively, in Figure 4.2, show that the original speech, has more robust energy content (in dB) than the waveforms of the received speeches. This is because during transmission by network equipment, the amplitude of the transmitted speech suffered attenuation thereby degrading the received speech signals.

### 4.7.2 Spectra Structures of Original and Received Speech Signals

Looking at the spectral structure of the original speech signal in Figure 4.3, it has a peak of -24 dB at about 50 Hz and diminished in intensity to -83 dB at a frequency of about 22,000 Hz. This is a very robust speech signal in the intensity spread over a bandwidth of 50 to 22,000 Hz. The spectra plots of the received speech signals from networks A, B, and C in Figures 4.4, 4.5 and 4.6 were band-limited within 86 to about 4,000 Hz with the network equipment. The band-limiting, which was done for the sake of bandwidth resource management, led to reduction in the robustness energy contents and spread of the received speech signals, and so partly responsible for degradations suffered by the received signals.

The effects of the band-limiting the transmitted speech by the networks and the energy spread, is also made glaring with the plot of the spectral analysis of the sample original speech and its corresponding received speeches in Figure 4.7.

Degradation of the received speeches caused by the shrinking of their energy contents and spread as a result of band-limiting of the transmitted speech signals have adverse effects on the intensity or magnitude of the resultant sensation of sound, which is loudness. All the psychoacoustic parameters of speech described in Section 2.11 are negatively affected, from loudness to sharpness, pitch and timbre, which is the tonal colouration of the speech signals, for they being frequency dependent or related, and the speech quality thereby degraded.

### 4.7.3 Programming of Loudness Estimations

Loudness was described in Sub-section 2.11.1 and was obtained with respect to the spectral density of the speech signals. As can be seen in Figures 4.3 to 4.7, the spectral densities of the received speech signals are not as robust as that of the original speech signal as a result to the band-limiting by network equipment.The

computerised estimation done for the loudness parameters of these speech signals helped quantify the shrinking of the spectral densities of the received signals as compared to that of the original speech signal.

The data on Tables 4.2 and 4.3 provide comparative analysis of the quantified loudness parameters for the original and the received speech signals. The maximum instantaneous loudness of the received speeches are: 42.55, 37.08 and 35.64% compared to that of the original speech, respectively.The maximum instantaneous loudness level of the received speeches are: 87.06, 84.98 and 84.37% compared to that of the original speech, respectively. This comparison provides a picture of the estimated quality of speechestransmitted over the telecommunication networks.

### 4.7.4 Results of Subjective Listening-only Tests

Results of the quality scores obtained from subjects for the degraded speeches obtained from the three Networks, A, B, and C in Section 4.4 were summarised as presented in Table 4.14. Network C has the highest variation of the quality scores with lowest quality score of 2.371 and maximum quality score of 3.595, while Network A has the lowest variation of the quality scores with lowest quality score of 2.522 and maximum quality score of 3.282.

### 4.7.5 Results of Intrusive Objective Quality Tests

Results of the quality scores obtained from the objective quality test using PESQ model for the degraded speeches obtained from the three Networks, A, B, and C in Section 4.5 were summarised as presented in Table 4.15. Network C has the highest variation of the quality scores and the lowest quality score is 1.963 and maximum quality score of 3.423, as was also observed for the results of the subjective quality tests. Network A has the lowest variation of the quality scores with the lowest quality score of 2.052 and maximum quality score of 3.178 as was also observed for subjective quality tests.

### 4.7.6 Results of Statistical Analysis

The regression lineswhich are least square regression lines, shown in Figures 4.7 to 4.9, provide the best fit to the scatter data points. The directions of the regression lines

Table 4.14 Summary of Subjective MOS Results for Received Speeches.

| Networks | MOS Mean ±Variance | Min MOS-LSQ | Max MOS-LQS |
|---|---|---|---|
| A | 2.902±0.380 | 2.522 | 3.282 |
| B | 2.952±0.447 | 2.505 | 3.399 |
| C | 2.983±0.612 | 2.371 | 3.595 |

Table 4.15 Summary of Intrusive Objective Quality Scores for Received Speeches

| Networks | MOS Mean ±Variance | Min MOS-LQO | Max MOS-LQO |
|---|---|---|---|
| A | 2.615±0.563 | 2.052 | 3.178 |
| B | 2.589±0.594 | 1.995 | 3.183 |
| C | 2.693±0.730 | 1.963 | 3.423 |

in each of the figures show that the mapped PESQ quality scores (PESQ MOS-LQO) are positively correlated with the Subjective MOS scores. The scatter plots provide

means by which given a subjective score for a particularly degraded speech, the real objective quality score could be predicted.

The correlation coefficients given in Table 4.6 are very strong, showing that the mapped objective quality scores have good closeness of fit with the subjective quality scores. The MOS-LQO correlates well with the estimated MOS-LQS, and this correlation is taken as a good figure of merit for the objective speech quality assessment (Dubey and Kumar, 2013).For Network A, it is 0.854, for Network B it is 0.871, while for Network C it is 0.848. Network B showed best correlation among the tested networks.

The assessment and correlation errors given on Table 4.8 as the RMSE for Networks A, B, and C, respectively, showed that the adopted method for this study was most efficient with Network A, which has the least RMSE. The prediction error values given on Table 4.9 showed that Network A also has the lowest prediction error, meaning that it can be most reliably predicted from the correlation of the PESQ MOS-LQO with the subjective MOS.

### 4.7.7 Results of Optimised Logistic Function Parameters

The obtained logistic mapping function shown in Figure 4.11 achieved 98.6% coverage of the range of the generic quality score, MOS. When compared with the existing two known international mapping functions, ITU-T Rec. P.862.1 and Morfitt III and Cotanis, as specified in Table 4.11 and shown in Figure 4.12, the optimised mapping function proved to have thebestquality score coverage over theirs. ITU-T Rec. P.862.1mapping function has 86.8% and Morfitt III and Cotanis mapping function has 93.7% coverage of the MOS range. This is an improvement of 11.8% over the coverage of ITU-T P.862.1 mapping function and 4.9% over thecoverage of Morfitt III and Cotanis mapping function.

### 4.7.8 Discussion of the ANOVA Test Results

From the test and analysis of three sets of mapped scores using the ITU-T Rec. P.862.1, the Morfitt III and Cotanis, and the developed function using hypothesis testing and Analysis Of Variance (ANOVA) at a significance level of $\alpha = 0.05$, gave results of F-statistical value of 60.6042, a critical-F of 3.04, and a $p$-value of 4.61721E-21. With $p< 0.05$, the Null Hypothesis was rejected, and the critical-F value being less than the F-statistic value confirmed the rejection. Therefore, the data

distribution of at least one of the functions has a different mean and belongs to a separate population of performance.

# CHAPTER FIVE

# CONCLUSION AND RECOMMENDATIONS

## 5.1 Conclusion

The paramount issue about provision of telecommunication services is the quality of experience gained by the subscribers.In ensuring that the quality of provisioned services that meet and surpass stipulated quality values, measurement, evaluation and reporting of QoS became paramount. This work as part of efforts at correcting the imbalance that existed in the approaches for assessment of the QoS of transmitted speeches over mobile telecommunication networks, being largely network-centric, addressed it from users' perceptual assessment perspective.

The speech database of 64 original and 192 received speeches was locally developed for this work. It allowed for naturalness and intelligibility of the speeches and also enhanced effectiveness of the speech quality testing techniques adopted for the study, particularly for the subjects that participated in the subjective testing.The received speeches were transmitted over three telecommunication Networks A, B, and C and became degraded by the characteristics of the networks.

The study of the psychoacoustic parameters of both the original and degraded speeches based on Zwicker's loudness model for the calculation of loudness parameters led to the development of a comparative speech quality assessment technique.This approach based on loudness parameter alone without undergoing perceptual transformation took into account the facts of the use of several noise reduction and suppression algorithms built into wireless telecommunication network operations to degrade or eliminate noise powers from received signals.

With the users' perceptual perspective,the focus was objective approach in which the PESQ model was used for E2E quality assessment, irrespective of the mobile network technology types and configurations. Correlating the mapped objective quality scores based on ITU-T Rec. P.862.1 mapping function with the subjective MOS values, the coefficient for degraded speeches from the three networks were obtained as 0.854, 0.871 and 0.848 respectively.

An improved logistic mapping function for mapping raw quality scores obtained from the use of PESQ model for objective testing of received speeches to the subjective MOS scores was developed for more appropriate scaling of quality scores. This was achieved with optimisation of the parameters of the steepness and factor of

the initial constant of integration of the logistic function. Upon comparative evaluation of the improved mapping function with two known international standard mapping functions, the ITU-T Rec. P.862.1 and the Morfitt III and Cotanis mapping functions, the developed function showed an improvement of 11.8 and 4.9 % of the coverage of the MOS scores range over those of these functions respectively.

## 5.2 Recommendations

PESQ has been around for a while, and its successor, POLQA, developed for next generation networks, is yet to be fully tested and developed on live networks across board. There is therefore need for more research efforts at enhancing the proficiencies of PESQ along with provingand fine-tuning the capabilities of POLQA, and the development of more innovative intrusive objective speech quality assessment techniques and algorithms. These would also require improved knowledge and use of highly proficient optimisation techniques.

Non-intrusive approach to objective assessment of the quality of coded and transmitted speech and other voice services on telecommunication networks, though highly computational, offers great promises for the development of very innovative speech quality estimation techniques. This is due to its use of only the output (degraded) speech and does not require that the original speech be present in the computation at the user's location.

Very little research efforts have been made and few quality estimation techniques and models have been so far developed on the non-intrusive approach. Research efforts at developing more and improved non-intrusive quality assessment techniques will requireintelligent and painstaking work and sufficient knowledge of machine learning techniques.It is highly recommended that future research efforts be more focused on the non-intrusive approach, knowing that reference speeches are in most cases not present with the receiving parties of voice calls.

# REFERENCES


Ahlbom, G, Bimbot, F. and Chollet, G., 1987. Modeling Spectral Speech Transitions Using Temporal Decomposition Techniques. *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '87),* Dallas, Texas, USA,6-9 April, 1987 12: 13 – 16.

Al-Mashouq, K., Aburas, A., and Maqbool, M. 2012. Speech Quality Assessment in Mobile Phones Using a Reduced-Complexity Algorithm. *Proceeding of the 2nd International Conference on Mobile Services, Resources, and Users (Mobility 2012), IARIA*, 2012 51-54.

Atal, B. S. 1983. Efficient coding of LPC parameters by temporal decomposition. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '83)*, Boston, Massachusetts, USA, 14-16 April, 1983 8: 81–84.

Atif, Y. and Zhang, L. 2014. Network resource management in support of QoS in ubiquitous learning.*Elsevier Journal of Network and Computer Applications.* May 2014 41: 48–156.

Audacity 2.1.2. 2015. Audacity is a trademark of Dominic Mazzoni. http://audacityteam.org.

Avertisyan, H. and Holub, J. 2018. Subjective speech quality measurement with and without parallel task: Laboratory test results comparison. *PLOS One*. July 2018. 13.7: 1 – 8. https://doi.org/10.1371/journal.pone.0199787.

Barriac, V., Le Saout, J.-Y., and Lockwood, C. 2004. Discussion on unified objective methodologies for the comparison of voice quality of narrowband and wideband scenarios.*Workshop on Wideband Speech Quality in Terminals and Networks: Assessment and Prediction*, *Mainz, Germany. Sponsored by France Telecom R & D.*8th and 9th June 2004

Baumgarte, F. 2002. Improved Audio Coding Using a Psychoacoustic Model Based on a Cochlear Filter Bank. *IEEE Transactions on Speech and Audio Processing*, October 2002 10.7: 495-503.

Bayya, A. and Vis, M. 1996. Objective Measures for Speech Quality Assessment in

Wireless Communications. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996, (ICASSP-96)*, Atlanta, Georgia, USA, 7-10 May, 1996 1: 495-498.

Bernasconi, M. and Seri, R. 2016. What are estimating when we fit Steven's power law? *ELSEVIER Journal of Mathematical Psychology.* December 2016. 75: 137 – 149.

Bonnans J. F., Gilbert, J. C., Lemarechal, C., and Sagastizabal, C. A. 2006. *Numerical Optimization – Theoretical and Practical Aspects.* Second Edition. Universitext. Springer-Verlag Berlin Heildelberg. New York. 23.

Bruce, I. C., Erfani, Y., and Zilany, M. S. A. 2018. A phenomenological model of the synapse between inner hair cell and auditory nerve: implications of limited neurotransmitter release sites. *ELSEVIER Journal of Hearing Research.* March 2018. 360: 40 – 54.

Chandra, S, Jayadeva, Mehra, A. 2009. *Numerical Optimisation with Applications*. Narosa Publishing House, Pvt.Ltd., 22 Delhi Medical Association Road, Daryaganj, New Delhi, 1100002, India. 2 – 9,

Chen, Z. and Hu, G. 2012. A revised method of calculating auditory excitation patterns and loudness for time-varying sounds. *In Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '12)*, *Kyoto, Japan*, 25-30 March, 2012. 157-160.

Comandur, B. K. 2011. Using the Levenberg Marquardt Algorithm forCamera Calibration without the Analytical Jacobian. *Robot Vision Laboratory, Purdue, West Lafayette, IN, USA.* bcomandu@purdue.edu.

Conway, A. E. 2004. Output-Based Method of Applying PESQ to Measure the Perceptual Quality of Framed Speech Signals. *In Proceedings of IEEE Wireless Communications and Networking Conference, 2004, WCNC '04, Atlanta, GA, U.S.A*, 21-25 May, 2004 4: 2521-2526.

Cotanis, I. 2009. The PESQ algorithm as the solution for speech quality evaluation on 2.5G and 3G networks. *Ascom Technical Paper.*

Cote, N., 2011. *Integral and Diagnostic Intrusive Prediction of Speech Quality*, Springer Book Series. Springer-Verlag Berlin Heidelberg. 2011. 37-85.

Deng, L. and O'Shaughnessy, D. 2003. *Speech Processing: a Dynamic and Optimization-Oriented Approach*. Marcel Dekker, Inc., USA–New-York, NY. 6. 74-132.

Dhkichi, S., Oukarfi, B., Fakkar, A. and Belbounaguia. 2014.Parameter identification of solar cell model using Levenberg–Marquardtalgorithm combined with simulated annealing. *Elsevier Journal of Solar Energy*.18 November, 2014 110: 781–788. http://dx.doi.org/10.1016/j.solener.2014.09.033

Dimolitsas, S., Corcoran, F. L. and Ravishankar, C. 1995. Dependence of Opinion Scores on Listening Sets Used in Degradation Category Assessments. *IEEE Transactions on Speech and Audio Processing*, September, 1995 3. 5: 421-424.

Dubey, R. K. and Kumar, A. 2013. Non-Intrusive Objective Speech Quality Assessment using a Combination of MFCC, PLP and LSF Features. *Proceedings of IEEE International Conference on Signal Processing and Communication (ICSC)*, December 2013 297-302.

Duc-Hung, L., Cong-Kha, P., Thrang, N. T. T., and Tu, B. T. 2012. Parameter Extraction and Optimization usingLevenberg-Marquardt algorithm. *In Proceeding of The Fourth International Conference on Communications and Electronics (ICCE)*, Hue, Vietnam 1-3 Aug. 2012. Published on IEEE Xplore on 02 October 2012 434 – 437.

Esmailpour, A and Nasser, N. 2011. Dynamic QoS-Based Bandwidth Allocation Framework for Broadband Wireless Networks. *IEEE Transactions on Vehicular Technology*, July 2011.

ETSI, 1999. "Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels," *ETSI EN 301 708 Recommendation*, 1999.

Falk, T. H., Xu, Q. and Chan, W. Y. 2005. Non-intrusive GMM-based speech quality measurement. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005, (ICASSP '05).* March 18-23, 2005 1: 125 - 128.

Falk, T. H. and Chan, W. Y. 2006a. Non-Intrusive Speech Quality Estimation Using Gaussian Mixture Models. *IEEE Signal Processing Letters,* February 2006 13.2: 108-111.

Falk, T. H. and Chan, W. Y. 2006b. Single-Ended Speech Quality Measurement Using Machine Learning Methods. *IEEE Transactions on Audio, Speech, and Language Processing*, November 2006 14. 6: 1935-1947.

Ferguson, J. and Brewster, S. A. 2017. Evaluation of Psychoacoustic Sound

Parameters for Sonification. *In Proceeding of 19th ACM International Conferenceon Multimodal Interaction* (ICMI 2017), Glasgow, Scotland, 13-17 Nov 2017 120-127. ISBN 9781450355438 (doi:10.1145/3136755.3136783)

Flanagan J. L. 1962. Computational Model for Basilar-Membrane Displacement. *The Journal of the Acoustic Society of America*, September 1962 34.8(2): 1370 – 1376.

Flanagan, J. L. 1972. *Speech Analysis Synthesis and Perception.* Second Edition. ringer-Verlag Berlin . Heidelberg New York.1972 9 – 15.

Freedman, D. S., Cohen, H. I., Deligeorges, S., Karl, C. and Hubbard, A. E. 2014. An Analog VLSI Implementation of the Inner Hair Cell and Auditory Nerve Using a Dual AGC Model. *IEEE Transactions on Biomedical Circuits and Systems*, April 2014 8.2: 240 – 256.

Gavin, H. P. 2019. The Levenberg-Marquardt method for nonlinear least squares curve-fitting problems. http://people.duke.edu/~hpgavin/ce281/lm.pdf. Retrieved August 22, 2019.

Genesis IB/RP/10003, 2009. History and description of loudness models. *GENESIS S. A.* December, 2009.

GENESIS S. A. 2009. Loudness Toolbox Matlab software. www.genesis.fr. December, 2009.

Ghitza, O. 1994. Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition. *IEEE Transactions On Speech And Audio Processing*, January 1994 2.1.11: 115 – 132.

Glasberg, B. R. and Moore, B. C. R. 2002. A Model of Loudness Applicable to Time-Varying Sounds. *Journal of Audio Engineering Society*, May 2002 50.5: 331-342.

Goldstein, T. and Rix, A. W. 2004. Perceptual Speech Quality Assessment in Acoustic and Binaural Applications. *In Proceedings of IEEE InternationalConference on Acoustics, Speech and Signal Processing, ICASSP'04*, Montreal, Quebec, Canada, 17-21 May, 2004 3: 1064-1067.

Grancharov, V., Zhao, D. Y., Lindblom, J., and Kleijn, W. B. 2006. Low-Complexity, Nonintrusive Speech Quality Assessment. *IEEE Transactions on Audio, Speech and Language Processing*, November 2006 14.6: 1948-195.

Grancharov, V. and Kleijn, W. B. 2008. *Springer Handbook of Speech Processing*,

Chapter on Speech Quality Assessment. Springer – Verlag, Berlin Heidelberg. 83–99.

Gray, P., Hollier, M. P., and Massara, R. E. 2000. Non-intrusive speech-quality assessment using vocal-tract models. *In Proceedings of Institute of Electrical Engineering (IEE), Vision, Image, Signal Process*, Dec. 2000 147.6: 493–501.

Hall, J. L. 2000. Application of multidimensional scaling to subjective evaluation of coded Speech. *In Proceedings of 2000 IEEE Workshop on Speech Coding,* Delavan, WI, 17-20 September, 2000 1:20-22.

Hauenstein, M. 1998. Application of Meddis' inner hair-cell model to the prediction of subjective speech-quality. *In Proceedings of 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 98,* Seattle, Washington, USA, May 1998 1: 545-548.

Hewitt, M. J. and Meddis, R. 1991. An evaluation of eight computer models of mammalian inner hair-cell function. *Journal of Acoustical Society of America*, August, 1991 90.2.1: 904 – 917.

Hiese, D. R. 1970. The Semantic Differential and Attitude Research. *Attitude Measurement*. Edited by Gene F. Summers. Chicago: Rand McNally, 1970 235-253.

Hines, A., Skoglund, J., Kokaram, A., and Harte, N. 2013. Robustness of speech quality metrics to background noise and network degradations: comparing VISQOL, PESQ and POLQA. *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '13.* Vancouver, BC, Canada. 26-31 May, 2013. 3697 – 3701.

Honda, K. 2008. Physiological Processes of Speech Production. Part A2. *Handbook of Speech Processing.* Edited by Benesty, J.; Sondi, M. M.; and Huang, Y. Springer–Verlag, Berlin Heindelsberg. 7-26.

Hou, J., Rabiner, L., and Dusan, S. 2006. Auditory Models for Speech Analysis. *Automatic Speech Attribute Transcription (ASAT) Projects*, Rutgers University, http://www.ece.gatech.edu/research/labs/asat/slides/meet-111204/Auditory-lrr.pdf, October 13, 2006.

Howard, D. M. and Angus, J. A. S. 2009. *Acoustics and Psychoacoustics*, 4th Edition, Elsivier Ltd., Focal Press, 30 Corporate Drive, Suite 400, Burlington, MA 01803, USA. 73-119.

Hu, Y. and Loizou, P. C. 2008. Evaluation of Objective Quality Measuresfor Speech

Enhancement. *IEEE Transactions On Audio, Speech, And Language Processing,* January 2008. 16. 1: 229 – 238.

Irino, T. and Unoki, M. 1998. A Time-Varying, Analysis/Synthesis Auditory Filterbank using the Gammachirp. *In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, Seattle, WA, USA, 12-15 May 1998 6: 3653 – 3656.

Irino, T. and Patterson, R. D. 2006. A Dynamic Compressive Gammachirp Auditory Filterbank. *IEEE Transactions on Audio, Speech and Language Processing*, November 2006 14.6: 2222–2232.

ITU-T Rec. E.800. 2008. *Definitions of terms related to quality of service.* ITU-T, Geneva, Switzerland, ITU-T Recommendation E.800, 09/2008.

ITU-T Rec. E.802. 2007. *Framework and Methodologies for the determination and application of QoS parameters*. ITU-T Geneva, Switzerland, ITU-T Recommendation, E.802, 02/2007.

ITU-T Rec. E.803. 2011. *Quality of service parameters for supporting service aspects*. ITU-T Geneva, Switzerland, ITU-T Recommendation E.803, 02/2011.

ITU-T Rec. G.729. 1996. *Coding of speech at 8 kbit/s using conjugate structure algebraic code-excited linear-prediction (CS-CELP) - Annex B: A silence compression scheme for G.729 optimized for terminals conforming to ITU-T Recommendation V.70*, ITU-T Geneva, Switzerland, ITU-T Recommendation G.729B, 11/1996.

ITU-T Rec. G.1000. 2001. *Communications quality of service: A framework and definitions*. ITU-T Geneva, Switzerland, ITU-T Recommendation G.1000, 11/2001.

ITU-T Rec. G.1010. 2001. *End – User Multimedia QoS categories*. ITU-T, Geneva, Switzerland, ITU-T Recommendation G.1010, 11/2001.

ITU-T Rec. P.10. 2006. *Vocabulary for performance and quality of service*. ITU-T Geneva, Switzerland, ITU-T Recommendation P.10, 07/2006.

ITU-T Rec. P.562. 2004. *Analysis and interpretation of INMD voice-service measurements*. ITU-T Geneva, Switzerland, ITU-T Recommendation P.562, 05/2004.

ITU-T Rec. P.563. 2004. *Single-ended method for objective speech quality assessment*

*in narrow-band telephony applications*. ITU-T Geneva, Switzerland, ITU-T Recommendation P.563, 05/2004.

ITU-T Rec. P.800. 1996. *Methods for Subjective determinationof Transmission Quality*. ITU-T Geneva, Switzerland, ITU-T Recommendation P.800, 08/1996.

ITU-T Rec. P.800.1. 2007. *Mean Opinion Score (MOS) terminology*. ITU-T Geneva, Switzerland, ITU-T Recommendation P.800.1, 07/2007.

ITU-T Rec. P.830. 1996. *Subjective Performance Assessment of Telephone Bad and Wideband Digital Codecs*. ITU-T Geneva, Switzerland, ITU-T Recommendation, P.830 02/1996.

ITU-T Rec. P.835. 2003. *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*. ITU-T Geneva, Switzerland, ITU-T Recommendation, P.835, 11/2003.

ITU-T Rec. P.861. 1996. *Objective quality measurement of telephone band (300 – 3400 Hz) speech codecs*. ITU-T Geneva, Switzerland, ITU-T Recommendation, P.861, 08/1996.

ITU-T Rec. P.862. 2001 *Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. ITU-T Geneva, Switzerland, ITU-T Recommendation, P.862, 02/2001.

ITU-T Rec. P.862.1 2003 *Mapping function for transforming P.862 raw result scores to MOS-LQO*. ITU-T Geneva, Switzerland, ITU-T Recommendation, P.862.1 11/2003.

ITU-T Rec. P. 862.2. 2007. *Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*. ITU-T Geneva, Switzerland,ITU-T Recommendation, P.862, 13 November, 2007.

ITU-T Rec. P.863. 2014. *Perceptual Objective Listening Quality Assessment (POLQA)*. ITU-T Geneva, Switzerland, ITU-T Recommendation, P.863 09/2014.

ITU-T Rec. BS.1387-1. 2001.*Method for objective measurements of perceived audio quality*. ITU-T Geneva, Switzerland, ITU-T Recommendation, BS.1387-1, 2001.

ITU-T P-Series Supplement 23, 1998. *ITU-T Coded-speech database*. ITU-T Geneva, Switzerland, ITU-T P-Series Supplement 23 02/1998.

James, R., Garside, J., Plana, L. A., Rowley, A. and Furber, S. B. 2018. Parallel distribution of an Inner Hair Cell and Auditory Nerve Model for Real-Time Application. *IEEE Transactions on Biomedical Circuits and Systems.* October 2018. 12.5: 1018 – 1026.

Ji, L-Q. 2013. Analysis of modified logistic model for describing the growth of durable customer goods in China. *Journal of Mathematical and Computational Applications*, 2013 18. 1: 30 – 37.

Jin, C. and Kubichek, R. 1996. Vector Quantization Techniques for Output-Based Objective Speech Quality. *Proceedings of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* June 1996 1: 491-494.

Jukic, D. and Scitovski, R. 2003. Solution of the least-squares problem for logistic function. *Elsevier Journal of Computational and Applied Mathematics*2003156: 159–177.

Jyothi, P. 2016. Basics of Speech Production. https://www.cse.iitb.ac.in/~pjyothi/cs753_spr16/slides/lecture4.pdf. Visited August 22, 2016.

Kajackas, A., Batkauskas, V., and Medeisis, A. 2004. Individual QoS Rating for Voice Services in Cellular Networks. *IEEE Communications Magazine*, June 2004 42.6: 88-93.

Kajackas, A. and Anskaitis, A. 2009. An Investigation of the Perceptual Value of Voice Frames. *INFORMATICA, 2009. Institute of Mathematics and Informatics, Vilnius*. 2009 20. 4: 487 – 498.

Kajackas, A and Vindasius, A. 2010. Analysis and Monitoring of End-user Perceived QoS in Mobile Networks. *14th IEEE Int'l Conference on Telecommunications Network Strategy & Planning Symposium, 2010, NETWORK '10,* Warsaw, 27-30 Sept 2010: 1-4.

Karjalainen, M. 1987. Auditory Models for Speech Processing, *In Proceeding of International Congress of Phonetic Sciences, ICPhS-87*, (Tallinn), 1987.

Kim, D.-S. and Tarraf, A. 2004. Perceptual Model for Non-Intrusive Speech Quality Assessment. *In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2004, (ICASSP-04),* Florence, Italy, 4-9 May, 2004 3:iii – 1060 – 3.

Kleczkowski, P. 1999. From ear modeling to auditory transform. *Archives of Acoustics*, 1999 2.24: 191-206.

Kola, J., Espy-Wilson, C., and Pruthi, T. 2011. Voice Activity Detection. *MERIT BIEN 2011 Final Report.*Retrieved from: http://www.ece.umd.edu/merit/archives/merit2011/merit_fair11_reports/report_Kola.pdfpp.1-6.

Kollmeier, B. 2008. Anatomy, Physiology, and Function of the Auditory System. *Handbook of Signal Processing in Acoustics*, Vol 1. Springer Science+Business Media LLC, 233 Spring Street, New York, NY 10013, USA. 147 – 158.

Kondo, K. 2012. *Subjective QualityMeasurement of SpeechIts Evaluation, Estimation and Applications*. Signals and Communication Technology Series. Springer-Verlag Berlin Heidelberg. http://www.springer.com/series/4748. DOI 10.1007/978-3-642-27506-7. 7 – 20.

Koster, F., Moller, S, Antons, J-N., Arndt, S., Guse, D., and Weiss, B. 2014. Methods for Assessing the Quality of Transmitted Speech and of Speech Communication Services. *Acoustics Australia*. December 2014 42. 3: 179 – 184.

Krishnamoorthi, H., Berisha, V., and Spanias A. 2008. A low-complexity loudness estimation Algorithm. *In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008,* Las Vegas, NV, March 31 2008-April 4 2008 361 – 364.

Kucharavy, D. and De Guio, R. 2015. Application of logistic growth curve. *Elsevier Procedia Engineering* 131 2015 280 – 290.

Kumar, R. and Saini, S. 2011. Measuring Parameters for speech quality in cellular networks. *International Journal of Advances in Computer Networks and its Security* 1: 746-194.

Levenberg, K., 1944. A Method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*. 2.2: 164–168.doi:10.1090/qam/10666 (https://doi.org/10.1090%2Fqam%2F10666).

Li, Y. and Jiang, L. 2013. Fitting Logistic Growth Curve with Nonlinear Mixed-effects Models. *Advance Journal of Food Science and Technology, Maxwell Scientific Publication Corp*. April 15, 2013 5: 392-397.

Lin, Y. and Abdulla, W. H. 2015. *Audio Watermark: A Comprehensive Foundation*

*Using MATLAB*. Springer International Publishing Co., Switzerland. 14 – 48.

Liu, W. M., Jellyman, K. A., Mason, J. S. D., and Evans, N. W. D. 2006. Assessment of Objective Quality Measures for Speech Intelligibility Estimation. *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP '06)*, Toulouse, 14-19 May, 2006 1: 1225 – 1228.

Lourakis, M. I. A. 2005. A Brief Description of the Levenberg-Marquardt Algorithm Implemented by levmar. *Institute of Computer ScienceFoundation for Research and Technology - Hellas (FORTH)*, Vassilika Vouton, P.O. Box 1385, GR 711 10, Heraklion, Crete, GREECE. February 11, 2005. Website: http://www.ics.forth.gr/~lourakis/levmar,Downloaded February 18, 2019.

Lyon, R. F. 1982. A Computational Model of Filtering, Detection, and Compression in the Cochlea. *In Proceeding of IEEE International Conference of Acoustics, Speech, and Signal Processing*, Paris, France, May 1982 1282 – 1286.

Lyon, R. F. 1984. Computational Models of Neural Auditory Processing. *In Proceeding of IEEE International Conference of Acoustics, Speech, and Signal Processing*, Sandiego, CA, March 1884 36.1: 1 – 4.

Lyon, R. F. 1986. Experiments with a Computational Model of the Cochlea. 1975-1978.

Lyon, R. F. and Mead, C. 1988. An Analog Electronic Cochlea. *IEEE Transactions on Acoustics, Speech, andSignal Processing.* July 1988 36.7: 1119 – 1134.

Lyon, R. F., Katsiamis, A. G., and Drakakis, E. M. 2010. History and Future of Auditory Filter Models. *In Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS),* Paris, May 30 2010-June 2, 2010 3809 – 3812.

Madsen, K., Nielsen, H. B., and Tingleff, O. 2004. *Methods for Non-linear Least Squares Problems*, 2nd Edition, April 2004. Informatics and Mathematical Modeling, Technical University of Denmark.

Mahdi, A. E. and Picovici, D. 2015. Enhanced output-based perceptual measure for predictingSubjective quality of speech. *13th European Signal Processing Conference, Antalya, Turkey.* 4-8 Sept. 2005. *Published by IEEE Xplore* https://ieeeexplore.ieee.org on 06 April, 2015.

Mahdi A. E. and Picovici D. 2006. Perceptual Voice Quality Measurement – Can You Hear Me Loud and Clear. *ICI Publishers*. 210–231.

Malfait, L, Gray, P., and Reed, M. J. 2008. Objective listening quality assessment of

speech communication Systems introducing continuously varying delay (time-warping): a Time alignment issue. *In Proceedings of 2008 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP '08).* Las Vegas, NV, USA.Date of Conference: 31 March - 4 April, 2008. Published 12 May, 2008. 4213-4216.

Marquardt, D. W. 1963. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, June1963 11.2: 431-441.

Mathworks. 2019. Lsqcurvefit. https://de.mathworks.com/help/optim/ug/lsqcurvefit.html?searchHighlight=lsqcurvefit&s_tid=doc_srchtitle. Website visited February 18, 2019.

Matthew J. A. 1992. BoundedPopulation Growth:A Curve Fitting Lesson. *Journal of Mathematics and Computer Education.* Spring 1992 26.2: 169-176.

Mayo, C., Clark, R. A. J., and King, S. 2005. Multidimensional scaling of listener responses to synthetic speech. *In Proceeding of Interspeech 2005*, Lisbon, Portugal, September 2005.

McEwan, A., and Schaik, A. 2002. A Silicon Representation of the Meddis Inner Hair Cell Model. *In Proceedings of the ICSC Symposia on Intelligent Systems & Application (ISA'2000),*Wollongong, Australia, December 2000 1544-078.

McEwan, A., and Schaik, A. 2003. An Analogue VLSI Implementation of the Meddis Inner Hair CellModel.*EURASIP Journal on Applied Signal Processing, Hindawi Publishing Corporation,* 2003 7: 639 – 648.

Meddis, R. 1986. Simulation of mechanical to neural transduction in the auditory receptor. *Journal of Acoustical Society of America,* March 1986 79.3: 702 – 711.

Meddis, R. and Lopez-Poveda, E. A. 2010. Auditory Peripheral: From Pinna to Auditory Nerve. *Computational Models of the Auditory System*. Edited by Ray Meddis, Enrique A. Lopez-Poveda, Arthur N. Popper, and Richard R. Fay. Springer Handbook of Auditory Research. Springer Science+Business Media, LLC, 233 Spring Street, New York, NY10013, USA 7 – 38.

Meng, M., Shang, W., Niu, W. and Gao, Q. 2014. A New Algorithm of Parameter Estimation forthe Logistic Equation in Modeling CO2 Emissions fromFossil Fuel Combustion. *Hindawi Publishing Corporation, Mathematical Problems*

*in Engineering.*2014, Article ID 616312, 5 pageshttp://dx.doi.org/10.1155/2014/616312

Miroslav, V. and Rozhon, J. 2012. Influence of Atmospheric Parameters on Speech Quality in GSM/UMTS. International Journal of Mathematical Models and Methods in Applied Sciences. 6. 4. 2012: 575 – 582.

Mishan, M. M., Zambello de Pinho, S., and Raquel de Carvalho, L. 2011. Determination of a point sufficiently close to the asymptotein nonlinear growth functions. Journal of Scientific Agriculture, Piracicaba, Brazilia.January/February 201168.1:109-114.

Mohamed, S. 2003. Perceptual Analysis Measurement System (PAMS). http://www.irisa.fr/armor/lesmembres/Mohamed/Thesis/node100.html

Moller, S. 2000. *Assessment and Prediction of Speech Quality in Telecommunications*, Springer Science + Business Media, Dordrecht, 2000.

Moore, B. C. J. 1987. Psychophysics of normal and impaired hearing. *British Medical Bulletin,* 1987 43.4: 887-908.

Moore, B. C. J., Glasberg, B. R. 1996. A Revision of Zwicker's Loudness Model, *Acta Acoustica*. 1996 82.

Moore, B. C. J., Glasberg, B. R., and Baer, T. 1997. A Model for the Prediction of Thresholds, Loudness, and Partial Loudness. *Journal of the Audio Engineering Society*, 45(4):224–240. 8, 34, 85, 216.

Moore, B. C. J. 1997. *An Introduction to the Psychology of Hearing*, 4th ed. London: Elsevier, Academic Press, 1997.

Moore, B. C. J. 2003. Coding of Sounds in the Auditory System and Its Relevance to Signal Processing and Coding in Cochlear Implants. *Otology & Neurotology,* 2003 24.2: 243-254.

More, J. J. and Wright, S. J. 1993. Optimisation Software Guide. *Society for Industrial and Applied Mathematics (SIAM)*. Argonne National Laboratory, Philadelphia. 7-19.

Morfitt III, J. C. and Cotanis, I. C. 2008. Mapping Objective Voice Quality Metrics to a MOS Domain for Field Measurements. United States Patent. Patent No. US007327985B2. *https://www.google.com/patents/US7327985* Feb 5, 2008.

Muñoz-Mulas, C.; Martínez-Olalla, R.; Gómez-Vilda, P.; Álvarez-Marquina, A.; and Mazaira-Fernández, L. M. 2013. Gender Detection in Running Speech from Glottal and Vocal Tract Correlates. *Proceedings of the 6th International*

*Conference on Nonlinear Speech Processing (NOLISP 2013)*, Mons, Belgium, June 19-21, 2013 25-32.

Necciari, T., Balazs, P., Holighaus, N., and Sondergaard, P. L. 2013. The ERBLET Transform: An Auditory-Based Time-Frequency Representation with Perfect Reconstruction. *In Proceeding of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '13),* Vancouver, BC, 26-31 May, 2013 498-502.

Ni, C., Gan. C., Li, W. and Chen, H. 2015. Bandwidth allocation based on priority andexcess-bandwidth-utilized algorithm in WDM/TDM PON. *Elsevier International Journal of Electronics and Communications*. November 201569.11: 1659-1666.

O'Shaughnessy, D. 2000. *Speech Communications – Human and Machine*. Second Edition. IEEE Press. 3 Park Avenue, 17th Floor, New York. 70-71.

OECD. 2006. Glossary of Statistical Terms. Last updated on Wednesday, March 8, 2006. https://stats.oecd.org/glossary/detail.asp?ID=5150

Olabisi, P. O. 2014. Trend Analysis of Key Cellular Network Quality Performance Metrics. *International Journal of Engineering Sciences & Research Technology*. July, 2014. 3. 7: 916-925.

Orovic, I and Stankovic, S. 2010. Time-Frequency-Based Characterisation and Eigenvalue Decomposition Applied to Speech Watermarking. EURASIP Journal in Signal Processing. 2010. 1 – 10. DOI: 10.1155/2010/572748.

Osgood, C. E., Tannenbaum, P. H., and Suci G. J. 1957. *The Measurement of Meaning.*Urbana: University of Illinois Press.

Ouadfeul, S-A. and Aliouane, L. 2015. Total Organic Carbon Prediction in Shale Gas Reservoirs fromWell Logs Data Using the Multilayer Perceptron Neural Networkwith Levenberg Marquardt Training Algorithm: Applicationto Barnett Shale. *Arab Journal of Science and Engineering. King Fahd University of Petroleum & Minerals.* Published online on 10 May, 2015.DOI 10.1007/s13369-015-1685-y

Oxenham, A. J. 2008. Pitch Perception and Auditory Stream Segregation: Implications for Hearing Loss and Cochlear Implants. *Trends in Amplification, SAGE Publications,* December 2008 12.4: 316-331.

Parrekh, J. 2010. 100 KPIs for Mobile Telecoms Operators, Entry on Consultant

Value Added, http://consultantvalueadded.com/2010/04/14/100-kpi%E2%80%99s-for-mobile-telecomoperators/

Patterson, R.D. and Holdsworth, J. 1991. A functional model of neural activity patterns and auditory images. *Advances in Speech, Hearing and Language Processing* , (W. A. Ainsworth, ed.), Vol 3. JAI Press, London.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand M. 1992. Complex Sounds and Auditory Images. *In Proceeding of 9th International Symposium on Hearing, Auditory physiology and perception,* Pergamon, Oxford. 1992 429-446.

Patterson, R. D., Allerhand, M. H., and Giguere, C. 1995 Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *Journal of Acoustical Society of America,* October, 1995 94.4: 1890 – 1894.

Pocta, P. and Beerends, J. G. 2015. Subjective and objective measurement of synthesised speech intelligibility in modern telephone conditions. *Elsevier Journal of Speech Communication*. 7. 2015: 1 – 9.

Psytechnics . 2004. PESQ: An Introduction. *Ipswich,* 23, Museum Street, Ipswich Suffolk, United Kingdom, IP1 1HN September, 2001 7.

Pulkki, V. and Karjalainen, M. 2015. *Communications Acoustics: An introduction to Speech, Audio and Psychoacoustics*. John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom. 111-131, 187.

Rabiner, L. R. and Juang, B-H. 1993. *Fundamentals of Speech Recognition.* PTR Prentice-Hall Inc, New Jersey 07632, 1993 17-20, 132.

Rabiner, L. R. and Schafer, R. W. 2007 Introduction to Digital Speech Processing. *Foundations and Trends in Signal Processing,*Netherlands, Drlft: Nov, 2007 1-2.

Ramírez, J., Górriz, J. M. and Segura, J. C. 2007. Voice Activity Detection. Fundamentals andSpeech Recognition System Robustness. Robust Speech Recognition and Understanding. Edited by Grimm, M. and Kroschel, K. I-Tech Education and Publishing, Vienna, Austria. June 2007 1 – 22.

Ranganathan, A., 2004. The Levenberg-Marquardt Algorithm. *The Semantic Scholar*.

8th June, 2004. https://www.semanticscholar.org/paper/The-Levenberg-Marquardt-Algorithm-Ranganathan/1e0078d36080d288dc240877fdb33f54ef5028c6..

Rennies, J., Verhey, J. L., and Fastl, H. 2010. *Comparison of loudness models for time-varying Sounds.* ACTA Acoustica United with America, 2010 9: 383-396.

Ritz, C.H., Burnett, I.S., and Lukasiak, J. 2000. Very low rate speech coding using temporal decomposition and waveform interpolation. *Proceedings of 2000 IEEE Workshop on Speech Coding,* 17-20 September 2000 29-31.

Rix, A. W and Hollier, M. P. 2000. The perceptual analysis measurement system for robust end-to-end speech assessment. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing – ICASSP*. 2000.1515-1518.

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. 2001. Perceptual Evaluation of Speech Quality (PESQ) – A New Method for Speech Quality Assessment of Telephone Networks and Codecs. *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01),* Salt Lake City, UT, 07-11 May, 2001 2: 749-752.

Rix, A. 2001. End – to – end speech quality assessment of networks using PESQ (P.862). *ITU-T SG12 Workshop*, 18-19 October, 2001.

Rix, A. W. 2004. Perceptual Speech Quality Assessment – A Review. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004, (ICASSP'04),* 3: 1056 – 1059.

Rix, A. W., Beerends, J. G., Kim, D. S., Kroon, P., and Ghitza, O. 2006. Objective Assessment of Speech and Audio Quality – Technology and Applications. *IEEE Transactions on Audio, Speech and Language Processing,* November 2006 14.6: 1890 – 1901.

Rohani, B.; Rohani, B.; Caldera, M. and Zepernick, H. –J. 2006. Benefits of perceptual speech quality metrics in modern cellular systems. *Institute of Engineering and Technology, Electronics Letters.* 12th October 2006 42 21.

Rozhon, J. and Voznak, M. 2011. Development of a speech quality monitoring tool based on ITU-T P.862. *34th International Conference on Telecommunications and Signal Processing*, Budapest, Hungary. 18-20 August, 2011. Published in *IEEE Xplore* on 13 October, 2011: 62-66.

Safuan H. M., Jovanoski, Z., Towers I. N., and Sidhu, H. S. 2013. Exact Solution of a non-autonomous logistic population model. *Ecological Modeling* (Journal of Elsevier), 251 (2013) 99-102. www.elsevier.com/locate/ecolmodel.

Sarabakha, A., Imanberdiyev, N., Kayacan, E., Khanesar, M. A. and Hagras, H. 2017. Novel Levenberg-Marquardt Based Learning Algorithm forUnmanned Aerial Vehicles. *Elsevier Journal of Information Sciences*. November 2017 417: 361 – 380. https://doi.org/10.1016/j.ins.2017.07.020

Schafer, M., Bahram, M., and Vary, P. 2013. An Extension of the PEAQ Measure by a Binaural Hearing Model. *In Proceedings of 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '13*, Vancouver, Canada, 26-31 May, 2013. 8164-8168.

Sechadri, G. and Yegnanarayana, B. 2009.Perceived loudness of speech based on the characteristics of glottal excitation source. *Journal of Acoustic Society of America.* 126 (4), October 2009. 2016 – 2017.

Seneff, S. 1986. A computational model for the Peripheral auditory system: Application to Speech recognition research. *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '86),* Tokyo, Apr 1986 11: 1983 – 1986.

Shiran, N. and Shallom, I. D. 2009. Enhanced PESQ algorithm for Objective assessment of Speech Quality at a continuous varying delay. *2009 International Workshop on Quality of Multimedia Experience,* San Diego, CA, USA. 29-31 July 2009. Published in IEEE Xplore on 18 September 2009: 157-162.

Simpson, A.J.R.; Terrell, M. J.; and Reiss J. D. 2013. A Practical Step-by-Step Guide to the Time-Varying Loudness Model of Moore, Glasberg and Baer (1997; 2002). Convention Paper 8873. *Audio Engineering Society 134th Convention*, Rome, Italy, May 4–7, 2013. 1 -7.

Skovenborg, E. and Nielsen, S. H. 2004.Evaluation of Different Loudness Models with Music and Speech Material. *Audio Engineering Society 117th Convention*, San Francisco, CA, USA. October 28–31, 2004 1 – 34.

Skrobacki, Z. 2007. Selected Methods for the estimationof the Logistic Function Parameters. *EKSPLOATACJA I NIEZAWODNOŚĆ NR* 3/2007 52 – 56.

Sloan, C., Harte, N., Kelly, D., Kokaram, A. C., and Hines, A. 2017. Objective

Assessment of Perceptual Audio Quality Using ViSQOLAudio. IEEE Transactions on Broadcasting. 63. 4. December, 2017: 693 – 705.

Smith, J. O. and Abel, J. S. 1999. Bark and ERB Bilinear Transforms. *IEEE Transactions on Speech and Audio Processing,* November 1999 7.6: 697-708.

Sohn, J.; Kim, N. S.; and Sung, W. 1999. A Statistical model-based voice activity detection*. IEEE Signal Processing Letters*. Jan. 1999 6. 1: 1-3.

Soliman, S. A. and Mantawy, A. H. 2012. Modern Optimization Techniques with Applications in ElectricPower Systems. *Springer Science+Business Media, LLC*, 233 Spring Street, New York,NY 10013, USA). 2012. DOI 10.1007/978-1-4614-1752-1 23–78.

Stefan, B., Ives, T., and Patterson, R. D. 2004. AIM-MAT: The Auditory Image Model in MATLAB. *Journal of ACTA – Acoustica United*, 2004 90: 781–787.

Stevens, K.N., Weismer, G. 2001.Acoustic phonetics.*The Journal of the Acoustical Society of America*. 2001 109: 17.

Summer, C. J., Lopez-Poveda, E. A., O'Mard, L. P., and Meddis, R. 2002. A revised model of the inner-hair cell and auditory-nerve complex. *Journal of Acoustical Society of America,* May 2002 111.5.1: 2178–2188.

Sun, L. and Ifeachor, E. C. 2006. Voice Quality Prediction Models and their Applications in VoIP Networks. *IEEE Transactions on Multimedia*, August 2006 8.4: 809-820.

Tsoularis, A. N. and Wallace, J. 2002. Analysis of Logistic Growth Model. *Journal of Mathematical Biosciences published by PubMed.*DOI: 10.1016/S0025-5564(02)00096-2. July 2002. 2: 23 – 46.

Villanueva, D. and Feijoo, A. E. 2016. Reformulation of parameters of the logistic function appliedto power curves of wind turbines. *Elsevier Journal of Electric Power Systems Research*. 137. 2016. 51–58 www.elsevier.com/locate/epsr

Voiers, W. 1977. Diagnostic acceptability measure for speech communication systems. *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1977, (ICASSP '77),* May, 1977 2: 204-207.

Volk, F. 2016. Predicting the loudness of non-stationary sounds: Zwicker's original envelope extraction vs. DIN 45631/A1: 2010. *Inter–Noise*, Hamburg, Germany. 2016. 1722 – 1728.

Voran, S. D. 1998. "Objective Estimation of Perceived Speech Quality Using Measuring Normalizing Blocks" *NTIA Report 98-347* April 1998.

Voran, S. 1999. Objective Estimation of Perceived Speech Quality—Part I: Development of the Measuring Normalizing Block Technique. *IEEE Transactions on Speech and Audio Processing,* July 1999 7.4: 371 – 382.

Voran, S. 1999. Objective Estimation of Perceived Speech Quality—Part II: Evaluation of the measuring normalizing block technique. *IEEE Transactions on Speech and Audio Processing,* July 1999 7.4: 385–390.

Voznak, M., Rozhon, J., Rezac, F., and Slachta, J. 2013. Real-Time Speech Quality Monitoring Using Non-Intrusive Method. *Journal of Recent Researches in Circuits, Communications and Signal Processing, World Scientific and Engineering Academy and Society (WSEAS) Publications*,201343-48.

Wang, S., Sekey, A., and Gersho, A. 1991. Auditory Distortion Measure for Speech Coding. *Proceedings of IEEE 1991 International Conference on Acoustics, Speech, and Signal Processing, (ICASSP 91),* Toronto, S., 1991 493-496.

Webster, P. and Jiricek, O. 2014. A Brief Comparison of Loudness Evaluation Models. *Akusticke listy*, 20(2), červenec 2014, str. 8–11.

Werner, M., Kumps, K., Tuisel, U., Beerend, J. G., and Vary, P. 2003. Parameter-Based Speech Quality Measures for GSM. *In Proceedings of The 14th IEEE 2003 International Symposium on Persona1,lndoor and Mobile Radio Communication, PIMRC 2003,* 7-10 Sept, 2003 3: 2611-2615.

Werner, M., Junge, T., and Vary, P. 2004. Quality Control for AMR Speech Channels in GSM Networks. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)* 3: 1076 – 1079.

Wikipedia, 2016. Window functions. https://en.wikipedia.org/wiki/window_function_(sql). Retrieved on 20 September, 2016.

Wikipedia, 2018. Levenberg-Marquardt Algorithm. https://en.wikipedia.org/index.php?title=Levenberg–Marquardt_algorithm&oldid=911154995 Retrieved 28 August 2018.

Yankayis, M. 1991. *Feature Extraction Mel-Frequency Cepstral Coefficients (MFCC).* Retrieved from https://www.ce.yildiz.edu.tr/personal/ fkarabiber/file/8172 / BLM5122_MFCC.pdf.

Ying, D.; Yan, Y.; Dang, J. and Soong, F. 2011. Voice Activity Detection Based On An Unsupervised Learning. *IEEE Transactions on Audio, Speech and Language Processing*, NOVEMBER 2011 19. 8: 2624 – 2633.

Zhang, X. 2005. Analysis of models for the synapse between the inner hair cell and the auditory nerve. *Journal of Acoustical Society of America,* September 2005 118.3.1: 1540 – 1553.

Zhang, W., Chang, Y., Liu, Y. and Xiao, L. 2013. A New Method of Objective Speech QualityAssessment in Communication System. *Journal of Multimedia,* June 2013 8. 3: 291-298.

Zhang, H, Li, D., Fu, X. and Bi, W. 2013. An improved Levenberg–Marquardt algorithm for extracting the features of Brillouin scatteringSpectrum. Measurement Science and Technology. IOP Publishing Ltd. 2013 015204 24: 1–5doi:10.1088/0957-0233/24/1/015204

Zhang, W., Chang, Y., Liu, Y. and Tian, Y. 2014. Performance analyze of QoE-based speech quality evaluation model. *In Proceedings of the 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Chengdu, China, 14-18 July 2014. Published on IEEE Xplore* on 08 September 2014. 1-6.

Zhang, Z. 2016. Mechanics of human voice production and control. *Journal of Acoustic Society of America*. October 2016. 140.4: 2614 – 2635.

Zilany, M. S. A., Bruce, I. A., and Carney, L. H. 2014. Updated parameters and expanded simulation options for a model of the auditory periphery. *Journal of Acoustic Society of America*. January 2014. 135.1: 283 – 286.

Zwicker, E., Flottorp, G., and Stevens, S. 1957. Critical Bandwidth in Loudness Summation. *Journal of the Acoustical Society of America*, 29(5):548–557. 8, 151, 152, 165, 216

Zwicker, E., and Fastl, H. 2007. Psychoacoustics: Facts and Models, 3rd Edition, *Springer-Verlag Berlin Heidelberg.*

**APPENDIX A:** Pseudo code for the Levenberg-Marquardt algorithm:

**Begin:**

**Input:** the objective function, the measurement vector, and the initial parameter values.

**Output:** Optimal parameter vector $\mathbf{p}^+ \in \mathbb{R}^m$ that minimise the non-linear least squares error, $\|\mathbf{x} - f(\mathbf{p})\|^2$.

$k$ := 0; $v$:= 2; **p**:= $\mathbf{p}_0$**;**

**A** := $\mathbf{J}^T$ **J**; $\epsilon_p$ := x $- f(p)$;**g** := $\mathbf{J}^T$ ***f***(**x**);

stop:= ($||\mathbf{g}||_\infty \leq \varepsilon_1$); $\mu$:= $\tau$ * $\max_{i=1,2,\dots,m}\{A_{ii}\}$**;**

while (not stop) and ($k<k_{max}$);

$k$ := $k$+1;

Repeat

Solve $(\mathbf{A} + \mu\mathbf{I})\delta_\mathbf{p} = -\mathbf{g}$;

if$||\delta_\mathbf{p}|| \leq \varepsilon_2 (||\mathbf{x}||+\varepsilon_2)$

stop:= true;

else

$\mathbf{p}_{new}$ := **p**+ $\delta_\mathbf{p}$;

$\rho$:= $(||\epsilon_p||^2 - ||\mathbf{x}-f(\mathbf{p}_{new})||^2) / (\delta_p^T (\mu\delta_\mathbf{p} + \mathbf{g}))$;

**if** $\rho$>0 {accepted step}

**p**:= $\mathbf{p}_{new}$

**A** := $\mathbf{J}^T$ **J**; $\epsilon_p$ := x $- f(p)$;**g** := $\mathbf{J}^T$ ***f***(**x**);

stop := ($||\mathbf{g}||_\infty \leq \varepsilon_1$) or ($||\epsilon_p||^2 \leq \varepsilon_3$)

$\mu$:=$\mu$∗ $\max\left(\frac{1}{3}, 1 - (2\rho - 1)^3\right)$; v := 2

**else**

$\mu$:=$\mu$∗$v$; $v$ := 2 ∗$v$

endif

endif

until ($\rho > 0$) or (stop)

endwhile

$\mathbf{p}^+$:= **p;**

(Ref. Loukaris, 2005).

**APPENDIX B: Original and Received Speech Files.**

Table B.1. Original Speech Files (in '.wav' format).

| 32 MALE SPEECH FILES | | | 32 FEMALE SPEECH FILES | | |
|---|---|---|---|---|---|
| **Speech file** | **Speech SPL ($dB_A$)** | **Room Noise ($dB_A$)** | **Speech file** | **Speech SPL ($dB_A$)** | **Room noise ($dB_A$)** |
| OM1S1 | 60 | 26 | OF1S1 | 56 | 29 |
| OM1S2 | 76 | 26 | OF1S2 | 59 | 28 |
| OM1S3 | 65 | 27 | OF1S3 | 60 | 26 |
| OM1S4 | 63 | 26 | OF1S4 | 64 | 26 |
| OM2S1 | 62 | 27 | OF2S1 | 58 | 26 |
| OM2S2 | 68 | 26 | OF2S2 | 48 | 28 |
| OM2S3 | 69 | 28 | OF2S3 | 56 | 29 |
| OM2S4 | 76 | 26 | OF2S4 | 55 | 26 |
| OM3S1 | 59 | 26 | OF3S1 | 59 | 29 |
| OM3S2 | 56 | 29 | OF3S2 | 62 | 28 |
| OM3S3 | 69 | 30 | OF3S3 | 64 | 29 |
| OM3S4 | 67 | 28 | OF3S4 | 69 | 27 |
| OM4S1 | 65 | 29 | OF4S1 | 60 | 28 |
| OM4S2 | 68 | 27 | OF4S2 | 64 | 26 |
| OM4S3 | 69 | 26 | OF4S3 | 65 | 27 |
| OM4S4 | 64 | 25 | OF4S4 | 63 | 29 |
| OM5S1 | 66 | 27 | OF5S1 | 49 | 28 |
| OM5S2 | 67 | 28 | OF5S2 | 53 | 29 |
| OM5S3 | 56 | 27 | OF5S3 | 57 | 27 |
| OM5S4 | 67 | 29 | OF5S4 | 54 | 26 |
| OM6S1 | 66 | 26 | OF6S1 | 59 | 28 |
| OM6S2 | 66 | 27 | OF6S2 | 60 | 29 |
| OM6S3 | 67 | 27 | OF6S3 | 62 | 29 |
| OM6S4 | 69 | 27 | OF6S4 | 57 | 30 |
| OM7S1 | 69 | 27 | OF7S1 | 59 | 30 |
| OM7S2 | 67 | 28 | OF7S2 | 65 | 30 |
| OM7S3 | 68 | 26 | OF7S3 | 62 | 30 |
| OM7S4 | 69 | 28 | OF7S4 | 51 | 28 |
| OM8S1 | 65 | 28 | OF8S1 | 69 | 28 |
| OM8S2 | 72 | 30 | OF8S2 | 67 | 28 |
| OM8S3 | 75 | 30 | OF8S3 | 69 | 28 |
| OM8S4 | 74 | 30 | OF8S4 | 68 | 29 |

Table B.2. Table of Received Speeches over Network A.

| Received Male Speech Files | Received Female Speech Files |
|---|---|
| AM1S1 | AF1S1 |
| AM1S2 | AF1S2 |
| AM1S3 | AF1S3 |
| AM1S4 | AF1S4 |
| AM2S1 | AF2S1 |
| AM2S2 | AF2S2 |
| AM2S3 | AF2S3 |
| AM2S4 | AF2S4 |
| AM3S1 | AF3S1 |
| AM3S2 | AF3S2 |
| AM3S3 | AF3S3 |
| AM3S4 | AF3S4 |
| AM4S1 | AF4S1 |
| AM4S2 | AF4S2 |
| AM4S3 | AF4S3 |
| AM4S4 | AF4S4 |
| AM5S1 | AF5S1 |
| AM5S2 | AF5S2 |
| AM5S3 | AF5S3 |
| AM5S4 | AF5S4 |
| AM6S1 | AF6S1 |
| AM6S2 | AF6S2 |
| AM6S3 | AF6S3 |
| AM6S4 | AF6S4 |
| AM7S1 | AF7S1 |
| AM7S2 | AF7S2 |
| AM7S3 | AF7S3 |
| AM7S4 | AF7S4 |
| AM8S1 | AF8S1 |
| AM8S2 | AF8S2 |
| AM8S3 | AF8S3 |
| AM8S4 | AF8S4 |

Table B.3. Table of Received Speeches over Network B.

| Received Male Speech Files | Received Female Speech Files |
|---|---|
| BM1S1 | BF1S1 |
| BM1S2 | BF1S2 |
| BM1S3 | BF1S3 |
| BM1S4 | BF1S4 |
| BM2S1 | BF2S1 |
| BM2S2 | BF2S2 |
| BM2S3 | BF2S3 |
| BM2S4 | BF2S4 |
| BM3S1 | BF3S1 |
| BM3S2 | BF3S2 |
| BM3S3 | BF3S3 |
| BM3S4 | BF3S4 |
| BM4S1 | BF4S1 |
| BM4S2 | BF4S2 |
| BM4S3 | BF4S3 |
| BM4S4 | BF4S4 |
| BM5S1 | BF5S1 |
| BM5S2 | BF5S2 |
| BM5S3 | BF5S3 |
| BM5S4 | BF5S4 |
| BM6S1 | BF6S1 |
| BM6S2 | BF6S2 |
| BM6S3 | BF6S3 |
| BM6S4 | BF6S4 |
| BM7S1 | BF7S1 |
| BM7S2 | BF7S2 |
| BM7S3 | BF7S3 |
| BM7S4 | BF7S4 |
| BM8S1 | BF8S1 |
| BM8S2 | BF8S2 |
| BM8S3 | BF8S3 |
| BM8S4 | BF8S4 |

Table B.4. Table of Received Speeches over Network C.

| Received Male Speech Files | Received Female Speech Files |
|---|---|
| CM1S1 | CF1S1 |
| CM1S2 | CF1S2 |
| CM1S3 | CF1S3 |
| CM1S4 | CF1S4 |
| CM2S1 | CF2S1 |
| CM2S2 | CF2S2 |
| CM2S3 | CF2S3 |
| CM2S4 | CF2S4 |
| CM3S1 | CF3S1 |
| CM3S2 | CF3S2 |
| CM3S3 | CF3S3 |
| CM3S4 | CF3S4 |
| CM4S1 | CF4S1 |
| CM4S2 | CF4S2 |
| CM4S3 | CF4S3 |
| CM4S4 | CF4S4 |
| CM5S1 | CF5S1 |
| CM5S2 | CF5S2 |
| CM5S3 | CF5S3 |
| CM5S4 | CF5S4 |
| CM6S1 | CF6S1 |
| CM6S2 | CF6S2 |
| CM6S3 | CF6S3 |
| CM6S4 | CF6S4 |
| CM7S1 | CF7S1 |
| CM7S2 | CF7S2 |
| CM7S3 | CF7S3 |
| CM7S4 | CF7S4 |
| CM8S1 | CF8S1 |
| CM8S2 | CF8S2 |
| CM8S3 | CF8S3 |
| CM8S4 | CF8S4 |

**APPENDIX C: Results of Subjective Test Scores.**

Table C.1. Subjective Test Scores for ReceivedSpeechesover Network A.

| Received Male Speech Files | Subjective MOS | Received Female Speech Files | Subjective MOS |
|---|---|---|---|
| AM1S1 | 3.5 | AF1S1 | 3.1 |
| AM1S2 | 3.0 | AF1S2 | 3.1 |
| AM1S3 | 2.7 | AF1S3 | 3.0 |
| AM1S4 | 3.5 | AF1S4 | 3.0 |
| AM2S1 | 2.6 | AF2S1 | 2.5 |
| AM2S2 | 2.7 | AF2S2 | 2.8 |
| AM2S3 | 2.5 | AF2S3 | 3.4 |
| AM2S4 | 3.1 | AF2S4 | 2.8 |
| AM3S1 | 3.0 | AF3S1 | 3.0 |
| AM3S2 | 3.4 | AF3S2 | 2.8 |
| AM3S3 | 3.0 | AF3S3 | 3.5 |
| AM3S4 | 3.4 | AF3S4 | 2.8 |
| AM4S1 | 2.8 | AF4S1 | 2.5 |
| AM4S2 | 3.2 | AF4S2 | 3.0 |
| AM4S3 | 3.3 | AF4S3 | 2.4 |
| AM4S4 | 3.1 | AF4S4 | 3.0 |
| AM5S1 | 2.5 | AF5S1 | 3.2 |
| AM5S2 | 2.4 | AF5S2 | 2.8 |
| AM5S3 | 2.1 | AF5S3 | 3.0 |
| AM5S4 | 2.5 | AF5S4 | 3.0 |
| AM6S1 | 3.2 | AF6S1 | 3.1 |
| AM6S2 | 3.0 | AF6S2 | 2.5 |
| AM6S3 | 3.0 | AF6S3 | 3.6 |
| AM6S4 | 2.6 | AF6S4 | 3.5 |
| AM7S1 | 3.0 | AF7S1 | 3.3 |
| AM7S2 | 2.9 | AF7S2 | 2.6 |
| AM7S3 | 2.5 | AF7S3 | 2.0 |
| AM7S4 | 2.8 | AF7S4 | 2.7 |
| AM8S1 | 2.7 | AF8S1 | 3.5 |
| AM8S2 | 2.5 | AF8S2 | 3.4 |
| AM8S3 | 2.5 | AF8S3 | 3.2 |
| AM8S4 | 2.0 | AF8S4 | 2.6 |

Table C.2. Subjective Test Scores for ReceivedSpeeches over Network B.

| Received Male Speech Files | Subjective MOS | Received Female Speech Files | Subjective MOS |
|---|---|---|---|
| BM1S1 | 2.6 | BF1S1 | 3.0 |
| BM1S2 | 3.0 | BF1S2 | 2.3 |
| BM1S3 | 2.5 | BF1S3 | 3.5 |
| BM1S4 | 3.0 | BF1S4 | 3.4 |
| BM2S1 | 2.5 | BF2S1 | 3.4 |
| BM2S2 | 3.5 | BF2S2 | 3.3 |
| BM2S3 | 2.8 | BF2S3 | 2.2 |
| BM2S4 | 2.4 | BF2S4 | 3.1 |
| BM3S1 | 2.8 | BF3S1 | 3.2 |
| BM3S2 | 3.0 | BF3S2 | 3.5 |
| BM3S3 | 3.0 | BF3S3 | 3.0 |
| BM3S4 | 3.5 | BF3S4 | 2.5 |
| BM4S1 | 3.0 | BF4S1 | 3.2 |
| BM4S2 | 3.5 | BF4S2 | 3.3 |
| BM4S3 | 3.3 | BF4S3 | 2.5 |
| BM4S4 | 3.0 | BF4S4 | 3.0 |
| BM5S1 | 2.5 | BF5S1 | 3.4 |
| BM5S2 | 2.0 | BF5S2 | 2.5 |
| BM5S3 | 2.4 | BF5S3 | 3.5 |
| BM5S4 | 2.5 | BF5S4 | 4.0 |
| BM6S1 | 3.0 | BF6S1 | 4.0 |
| BM6S2 | 3.0 | BF6S2 | 2.5 |
| BM6S3 | 2.5 | BF6S3 | 3.4 |
| BM6S4 | 3.2 | BF6S4 | 3.5 |
| BM7S1 | 3.0 | BF7S1 | 3.5 |
| BM7S2 | 2.5 | BF7S2 | 2.1 |
| BM7S3 | 3.0 | BF7S3 | 2.5 |
| BM7S4 | 2.8 | BF7S4 | 3.0 |
| BM8S1 | 3.0 | BF8S1 | 3.0 |
| BM8S2 | 2.6 | BF8S2 | 2.5 |
| BM8S3 | 2.5 | BF8S3 | 2.5 |
| BM8S4 | 3.2 | BF8S4 | 3.5 |

Table C.3. Subjective Test Scores for ReceivedSpeeches over Network C.

| Received Male Speech Files | Subjective MOS | Received Female Speech Files | Subjective MOS |
|---|---|---|---|
| CM1S1 | 4.0 | CF1S1 | 3.2 |
| CM1S2 | 3.5 | CF1S2 | 2.2 |
| CM1S3 | 2.8 | CF1S3 | 3.2 |
| CM1S4 | 3.3 | CF1S4 | 3.4 |
| CM2S1 | 2.6 | CF2S1 | 3.5 |
| CM2S2 | 2.5 | CF2S2 | 2.7 |
| CM2S3 | 2.8 | CF2S3 | 3.5 |
| CM2S4 | 2.0 | CF2S4 | 2.9 |
| CM3S1 | 2.9 | CF3S1 | 4.0 |
| CM3S2 | 3.2 | CF3S2 | 2.5 |
| CM3S3 | 3.2 | CF3S3 | 3.1 |
| CM3S4 | 4.2 | CF3S4 | 2.8 |
| CM4S1 | 2.2 | CF4S1 | 3.0 |
| CM4S2 | 2.3 | CF4S2 | 3.5 |
| CM4S3 | 3.6 | CF4S3 | 2.0 |
| CM4S4 | 2.6 | CF4S4 | 3.8 |
| CM5S1 | 1.6 | CF5S1 | 2.2 |
| CM5S2 | 1.9 | CF5S2 | 3.0 |
| CM5S3 | 3.6 | CF5S3 | 3.0 |
| CM5S4 | 2.5 | CF5S4 | 3.5 |
| CM6S1 | 2.9 | CF6S1 | 3.5 |
| CM6S2 | 3.8 | CF6S2 | 3.8 |
| CM6S3 | 3.0 | CF6S3 | 3.8 |
| CM6S4 | 3.2 | CF6S4 | 3.7 |
| CM7S1 | 2.6 | CF7S1 | 3.8 |
| CM7S2 | 2.2 | CF7S2 | 3.0 |
| CM7S3 | 3.5 | CF7S3 | 3.0 |
| CM7S4 | 3.8 | CF7S4 | 3.5 |
| CM8S1 | 2.5 | CF8S1 | 3.0 |
| CM8S2 | 2.1 | CF8S2 | 2.3 |
| CM8S3 | 2.5 | CF8S3 | 2.0 |
| CM8S4 | 2.3 | CF8S4 | 2.8 |

**APPENDIX D:Results of Raw PESQ Quality Test Scores.**

Table D.1. Results of Raw PESQ Quality Test Scores for Network A.

| Original Male Speech Files | Received Male Speech Files | PESQ Raw Quality Score | Original Female Speech Files | Received Female Speech Files | PESQ Raw Quality Score |
|---|---|---|---|---|---|
| OM1S1 | AM1S1 | 3.581 | OF1S1 | AF1S1 | 3.205 |
| OM1S2 | AM1S2 | 3.237 | OF1S2 | AF1S2 | 2.166 |
| OM1S3 | AM1S3 | 2.635 | OF1S3 | AF1S3 | 3.107 |
| OM1S4 | AM1S4 | 3.534 | OF1S4 | AF1S4 | 3.296 |
| OM2S1 | AM2S1 | 2.633 | OF2S1 | AF2S1 | 2.416 |
| OM2S2 | AM2S2 | 2.555 | OF2S2 | AF2S2 | 2.687 |
| OM2S3 | AM2S3 | 2.542 | OF2S3 | AF2S3 | 2.828 |
| OM2S4 | AM2S4 | 3.131 | OF2S4 | AF2S4 | 2.718 |
| OM3S1 | AM3S1 | 2.414 | OF3S1 | AF3S1 | 3.051 |
| OM3S2 | AM3S2 | 3.217 | OF3S2 | AF3S2 | 2.617 |
| OM3S3 | AM3S3 | 3.126 | OF3S3 | AF3S3 | 3.513 |
| OM3S4 | AM3S4 | 3.305 | OF3S4 | AF3S4 | 2.685 |
| OM4S1 | AM4S1 | 2.646 | OF4S1 | AF4S1 | 2.276 |
| OM4S2 | AM4S2 | 3.001 | OF4S2 | AF4S2 | 2.805 |
| OM4S3 | AM4S3 | 3.173 | OF4S3 | AF4S3 | 2.515 |
| OM4S4 | AM4S4 | 2.956 | OF4S4 | AF4S4 | 3.041 |
| OM5S1 | AM5S1 | 2.666 | OF5S1 | AF5S1 | 3.144 |
| OM5S2 | AM5S2 | 2.156 | OF5S2 | AF5S2 | 2.679 |
| OM5S3 | AM5S3 | 1.857 | OF5S3 | AF5S3 | 2.825 |
| OM5S4 | AM5S4 | 2.396 | OF5S4 | AF5S4 | 2.718 |
| OM6S1 | AM6S1 | 3.050 | OF6S1 | AF6S1 | 3.251 |
| OM6S2 | AM6S2 | 2.853 | OF6S2 | AF6S2 | 2.663 |
| OM6S3 | AM6S3 | 2.619 | OF6S3 | AF6S3 | 3.706 |
| OM6S4 | AM6S4 | 2.568 | OF6S4 | AF6S4 | 3.374 |
| OM7S1 | AM7S1 | 2.680 | OF7S1 | AF7S1 | 3.547 |
| OM7S2 | AM7S2 | 2.508 | OF7S2 | AF7S2 | 2.811 |
| OM7S3 | AM7S3 | 2.579 | OF7S3 | AF7S3 | 2.061 |
| OM7S4 | AM7S4 | 3.099 | OF7S4 | AF7S4 | 2.683 |
| OM8S1 | AM8S1 | 2.514 | OF8S1 | AF8S1 | 3.612 |
| OM8S2 | AM8S2 | 2.429 | OF8S2 | AF8S2 | 3.301 |
| OM8S3 | AM8S3 | 2.360 | OF8S3 | AF8S3 | 3.293 |
| OM8S4 | AM8S4 | 2.150 | OF8S4 | AF8S4 | 2.718 |

Table D.2. Results of Raw PESQ Quality Test Scores for Network B.

| Original Male Speech Files | Received Male Speech Files | PESQ Raw Quality Score | Original Female Speech Files | Received Female Speech Files | PESQ Raw Quality Score |
|---|---|---|---|---|---|
| OM1S1 | BM1S1 | 2.745 | OF1S1 | BF1S1 | 2.851 |
| OM1S2 | BM1S2 | 3.066 | OF1S2 | BF1S2 | 2.095 |
| OM1S3 | BM1S3 | 2.752 | OF1S3 | BF1S3 | 3.353 |
| OM1S4 | BM1S4 | 3.286 | OF1S4 | BF1S4 | 3.025 |
| OM2S1 | BM2S1 | 2.507 | OF2S1 | BF2S1 | 3.145 |
| OM2S2 | BM2S2 | 3.145 | OF2S2 | BF2S2 | 3.268 |
| OM2S3 | BM2S3 | 2.538 | OF2S3 | BF2S3 | 2.447 |
| OM2S4 | BM2S4 | 2.148 | OF2S4 | BF2S4 | 2.884 |
| OM3S1 | BM3S1 | 2.551 | OF3S1 | BF3S1 | 3.066 |
| OM3S2 | BM3S2 | 2.702 | OF3S2 | BF3S2 | 3.143 |
| OM3S3 | BM3S3 | 2.794 | OF3S3 | BF3S3 | 2.774 |
| OM3S4 | BM3S4 | 3.293 | OF3S4 | BF3S4 | 2.582 |
| OM4S1 | BM4S1 | 2.610 | OF4S1 | BF4S1 | 2.917 |
| OM4S2 | BM4S2 | 3.139 | OF4S2 | BF4S2 | 3.510 |
| OM4S3 | BM4S3 | 3.051 | OF4S3 | BF4S3 | 2.407 |
| OM4S4 | BM4S4 | 2.616 | OF4S4 | BF4S4 | 2.870 |
| OM5S1 | BM5S1 | 1.849 | OF5S1 | BF5S1 | 3.597 |
| OM5S2 | BM5S2 | 1.895 | OF5S2 | BF5S2 | 2.469 |
| OM5S3 | BM5S3 | 2.406 | OF5S3 | BF5S3 | 3.179 |
| OM5S4 | BM5S4 | 1.951 | OF5S4 | BF5S4 | 3.728 |
| OM6S1 | BM6S1 | 2.330 | OF6S1 | BF6S1 | 3.443 |
| OM6S2 | BM6S2 | 3.010 | OF6S2 | BF6S2 | 2.401 |
| OM6S3 | BM6S3 | 2.379 | OF6S3 | BF6S3 | 3.299 |
| OM6S4 | BM6S4 | 2.865 | OF6S4 | BF6S4 | 3.139 |
| OM7S1 | BM7S1 | 2.793 | OF7S1 | BF7S1 | 3.343 |
| OM7S2 | BM7S2 | 2.624 | OF7S2 | BF7S2 | 2.009 |
| OM7S3 | BM7S3 | 3.103 | OF7S3 | BF7S3 | 2.092 |
| OM7S4 | BM7S4 | 2.684 | OF7S4 | BF7S4 | 3.389 |
| OM8S1 | BM8S1 | 2.825 | OF8S1 | BF8S1 | 3.006 |
| OM8S2 | BM8S2 | 2.728 | OF8S2 | BF8S2 | 2.379 |
| OM8S3 | BM8S3 | 2.127 | OF8S3 | BF8S3 | 2.638 |
| OM8S4 | BM8S4 | 3.550 | OF8S4 | BF8S4 | 3.235 |

Table D.3. Results of Raw PESQ Quality Test Scores for Network C.

| Original Male Speech Files | Received Male Speech Files | PESQ Raw Quality Score | Original Female Speech Files | Received Female Speech Files | PESQ Raw Quality Score |
|---|---|---|---|---|---|
| OM1S1 | CM1S1 | 3.527 | OF1S1 | CF1S1 | 3.064 |
| OM1S2 | CM1S2 | 3.399 | OF1S2 | CF1S2 | 2.164 |
| OM1S3 | CM1S3 | 2.451 | OF1S3 | CF1S3 | 3.024 |
| OM1S4 | CM1S4 | 2.964 | OF1S4 | CF1S4 | 3.349 |
| OM2S1 | CM2S1 | 2.735 | OF2S1 | CF2S1 | 2.510 |
| OM2S2 | CM2S2 | 2.286 | OF2S2 | CF2S2 | 2.467 |
| OM2S3 | CM2S3 | 3.024 | OF2S3 | CF2S3 | 3.306 |
| OM2S4 | CM2S4 | 2.874 | OF2S4 | CF2S4 | 2.761 |
| OM3S1 | CM3S1 | 2.607 | OF3S1 | CF3S1 | 3.818 |
| OM3S2 | CM3S2 | 3.461 | OF3S2 | CF3S2 | 2.225 |
| OM3S3 | CM3S3 | 3.545 | OF3S3 | CF3S3 | 3.065 |
| OM3S4 | CM3S4 | 3.773 | OF3S4 | CF3S4 | 2.440 |
| OM4S1 | CM4S1 | 1.722 | OF4S1 | CF4S1 | 2.770 |
| OM4S2 | CM4S2 | 2.321 | OF4S2 | CF4S2 | 3.461 |
| OM4S3 | CM4S3 | 3.450 | OF4S3 | CF4S3 | 2.407 |
| OM4S4 | CM4S4 | 2.607 | OF4S4 | CF4S4 | 3.680 |
| OM5S1 | CM5S1 | 2.077 | OF5S1 | CF5S1 | 2.098 |
| OM5S2 | CM5S2 | 1.903 | OF5S2 | CF5S2 | 2.983 |
| OM5S3 | CM5S3 | 3.014 | OF5S3 | CF5S3 | 2.913 |
| OM5S4 | CM5S4 | 2.080 | OF5S4 | CF5S4 | 3.028 |
| OM6S1 | CM6S1 | 1.133 | OF6S1 | CF6S1 | 3.553 |
| OM6S2 | CM6S2 | 3.582 | OF6S2 | CF6S2 | 3.730 |
| OM6S3 | CM6S3 | 3.188 | OF6S3 | CF6S3 | 3.530 |
| OM6S4 | CM6S4 | 2.565 | OF6S4 | CF6S4 | 3.469 |
| OM7S1 | CM7S1 | 2.558 | OF7S1 | CF7S1 | 3.678 |
| OM7S2 | CM7S2 | 2.714 | OF7S2 | CF7S2 | 2.754 |
| OM7S3 | CM7S3 | 3.588 | OF7S3 | CF7S3 | 3.115 |
| OM7S4 | CM7S4 | 3.472 | OF7S4 | CF7S4 | 3.405 |
| OM8S1 | CM8S1 | 2.676 | OF8S1 | CF8S1 | 3.013 |
| OM8S2 | CM8S2 | 2.263 | OF8S2 | CF8S2 | 2.152 |
| OM8S3 | CM8S3 | 2.623 | OF8S3 | CF8S3 | 2.389 |
| OM8S4 | CM8S4 | 2.695 | OF8S4 | CF8S4 | 2.529 |

**APPENDIX E: Results of the Mapped PESQ Quality Scores.**

Table E.1. Results of the Mapped PESQ Quality Scores for Network A.

| Original Male Speech Files | Received Male Speech Files | PESQ MOS-LQO Score | Original Female Speech Files | Received Female Speech Files | PESQ MOS-LQO Score |
|---|---|---|---|---|---|
| OM1S1 | AM1S1 | 3.664 | OF1S1 | AF1S1 | 3.128 |
| OM1S2 | AM1S2 | 3.176 | OF1S2 | AF1S2 | 1.775 |
| OM1S3 | AM1S3 | 2.306 | OF1S3 | AF1S3 | 2.982 |
| OM1S4 | AM1S4 | 3.601 | OF1S4 | AF1S4 | 3.263 |
| OM2S1 | AM2S1 | 2.304 | OF2S1 | AF2S1 | 2.036 |
| OM2S2 | AM2S2 | 2.203 | OF2S2 | AF2S2 | 2.375 |
| OM2S3 | AM2S3 | 2.187 | OF2S3 | AF2S3 | 2.571 |
| OM2S4 | AM2S4 | 3.018 | OF2S4 | AF2S4 | 2.418 |
| OM3S1 | AM3S1 | 2.034 | OF3S1 | AF3S1 | 2.898 |
| OM3S2 | AM3S2 | 3.146 | OF3S2 | AF3S2 | 2.283 |
| OM3S3 | AM3S3 | 3.010 | OF3S3 | AF3S3 | 3.572 |
| OM3S4 | AM3S4 | 3.276 | OF3S4 | AF3S4 | 2.373 |
| OM4S1 | AM4S1 | 2.321 | OF4S1 | AF4S1 | 1.883 |
| OM4S2 | AM4S2 | 2.824 | OF4S2 | AF4S2 | 2.539 |
| OM4S3 | AM4S3 | 3.080 | OF4S3 | AF4S3 | 2.154 |
| OM4S4 | AM4S4 | 2.757 | OF4S4 | AF4S4 | 2.883 |
| OM5S1 | AM5S1 | 2.347 | OF5S1 | AF5S1 | 3.037 |
| OM5S2 | AM5S2 | 1.766 | OF5S2 | AF5S2 | 2.365 |
| OM5S3 | AM5S3 | 1.526 | OF5S3 | AF5S3 | 2.567 |
| OM5S4 | AM5S4 | 2.013 | OF5S4 | AF5S4 | 2.418 |
| OM6S1 | AM6S1 | 2.897 | OF6S1 | AF6S1 | 3.196 |
| OM6S2 | AM6S2 | 2.607 | OF6S2 | AF6S2 | 2.343 |
| OM6S3 | AM6S3 | 2.285 | OF6S3 | AF6S3 | 3.825 |
| OM6S4 | AM6S4 | 2.220 | OF6S4 | AF6S4 | 3.376 |
| OM7S1 | AM7S1 | 2.366 | OF7S1 | AF7S1 | 3.618 |
| OM7S2 | AM7S2 | 2.145 | OF7S2 | AF7S2 | 2.547 |
| OM7S3 | AM7S3 | 2.234 | OF7S3 | AF7S3 | 1.682 |
| OM7S4 | AM7S4 | 2.970 | OF7S4 | AF7S4 | 2.370 |
| OM8S1 | AM8S1 | 2.152 | OF8S1 | AF8S1 | 3.705 |
| OM8S2 | AM8S2 | 2.051 | OF8S2 | AF8S2 | 3.270 |
| OM8S3 | AM8S3 | 1.973 | OF8S3 | AF8S3 | 3.258 |
| OM8S4 | AM8S4 | 1.761 | OF8S4 | AF8S4 | 2.418 |

Table E.2. Results of the Mapped PESQ Quality Scores for Network B.

| Original Male Speech Files | Received Male Speech Files | PESQ MOS-LQO Score | Original Female Speech Files | Received Female Speech Files | PESQ MOS-LQO Score |
|---|---|---|---|---|---|
| OM1S1 | BM1S1 | 2.455 | OF1S1 | BF1S1 | 2.604 |
| OM1S2 | BM1S2 | 2.920 | OF1S2 | BF1S2 | 1.711 |
| OM1S3 | BM1S3 | 2.464 | OF1S3 | BF1S3 | 3.346 |
| OM1S4 | BM1S4 | 3.248 | OF1S4 | BF1S4 | 2.859 |
| OM2S1 | BM2S1 | 2.144 | OF2S1 | BF2S1 | 3.038 |
| OM2S2 | BM2S2 | 3.038 | OF2S2 | BF2S2 | 3.221 |
| OM2S3 | BM2S3 | 2.182 | OF2S3 | BF2S3 | 2.072 |
| OM2S4 | BM2S4 | 1.759 | OF2S4 | BF2S4 | 2.652 |
| OM3S1 | BM3S1 | 2.198 | OF3S1 | BF3S1 | 2.920 |
| OM3S2 | BM3S2 | 2.396 | OF3S2 | BF3S2 | 3.036 |
| OM3S3 | BM3S3 | 2.523 | OF3S3 | BF3S3 | 2.495 |
| OM3S4 | BM3S4 | 3.258 | OF3S4 | BF3S4 | 2.237 |
| OM4S1 | BM4S1 | 2.274 | OF4S1 | BF4S1 | 2.700 |
| OM4S2 | BM4S2 | 3.030 | OF4S2 | BF4S2 | 3.568 |
| OM4S3 | BM4S3 | 2.898 | OF4S3 | BF4S3 | 2.026 |
| OM4S4 | BM4S4 | 2.281 | OF4S4 | BF4S4 | 2.632 |
| OM5S1 | BM5S1 | 1.521 | OF5S1 | BF5S1 | 3.685 |
| OM5S2 | BM5S2 | 1.553 | OF5S2 | BF5S2 | 2.098 |
| OM5S3 | BM5S3 | 2.024 | OF5S3 | BF5S3 | 3.089 |
| OM5S4 | BM5S4 | 1.594 | OF5S4 | BF5S4 | 3.852 |
| OM6S1 | BM6S1 | 1.940 | OF6S1 | BF6S1 | 3.475 |
| OM6S2 | BM6S2 | 2.837 | OF6S2 | BF6S2 | 2.019 |
| OM6S3 | BM6S3 | 1.994 | OF6S3 | BF6S3 | 3.267 |
| OM6S4 | BM6S4 | 2.625 | OF6S4 | BF6S4 | 3.030 |
| OM7S1 | BM7S1 | 2.522 | OF7S1 | BF7S1 | 3.331 |
| OM7S2 | BM7S2 | 2.292 | OF7S2 | BF7S2 | 1.639 |
| OM7S3 | BM7S3 | 2.976 | OF7S3 | BF7S3 | 1.709 |
| OM7S4 | BM7S4 | 2.371 | OF7S4 | BF7S4 | 3.398 |
| OM8S1 | BM8S1 | 2.567 | OF8S1 | BF8S1 | 2.831 |
| OM8S2 | BM8S2 | 2.431 | OF8S2 | BF8S2 | 1.994 |
| OM8S3 | BM8S3 | 1.740 | OF8S3 | BF8S3 | 2.310 |
| OM8S4 | BM8S4 | 3.622 | OF8S4 | BF8S4 | 3.173 |

Table E.3. Results of the Mapped PESQ Quality Scores for Network C.

| Original Male Speech Files | Received Male Speech Files | PESQ MOS-LQO Score | Original Female Speech Files | Received Female Speech Files | PESQ MOS-LQO Score |
|---|---|---|---|---|---|
| OM1S1 | CM1S1 | 3.591 | OF1S1 | CF1S1 | 2.917 |
| OM1S2 | CM1S2 | 3.412 | OF1S2 | CF1S2 | 1.774 |
| OM1S3 | CM1S3 | 2.077 | OF1S3 | CF1S3 | 2.858 |
| OM1S4 | CM1S4 | 2.769 | OF1S4 | CF1S4 | 3.340 |
| OM2S1 | CM2S1 | 2.441 | OF2S1 | CF2S1 | 2.147 |
| OM2S2 | CM2S2 | 1.894 | OF2S2 | CF2S2 | 2.096 |
| OM2S3 | CM2S3 | 2.858 | OF2S3 | CF2S3 | 3.277 |
| OM2S4 | CM2S4 | 2.638 | OF2S4 | CF2S4 | 2.477 |
| OM3S1 | CM3S1 | 2.270 | OF3S1 | CF3S1 | 3.958 |
| OM3S2 | CM3S2 | 3.500 | OF3S2 | CF3S2 | 1.832 |
| OM3S3 | CM3S3 | 3.616 | OF3S3 | CF3S3 | 2.919 |
| OM3S4 | CM3S4 | 3.906 | OF3S4 | CF3S4 | 2.064 |
| OM4S1 | CM4S1 | 1.440 | OF4S1 | CF4S1 | 2.490 |
| OM4S2 | CM4S2 | 1.931 | OF4S2 | CF4S2 | 3.500 |
| OM4S3 | CM4S3 | 3.484 | OF4S3 | CF4S3 | 2.026 |
| OM4S4 | CM4S4 | 2.270 | OF4S4 | CF4S4 | 3.792 |
| OM5S1 | CM5S1 | 1.696 | OF5S1 | CF5S1 | 1.714 |
| OM5S2 | CM5S2 | 1.558 | OF5S2 | CF5S2 | 2.797 |
| OM5S3 | CM5S3 | 2.843 | OF5S3 | CF5S3 | 2.694 |
| OM5S4 | CM5S4 | 1.698 | OF5S4 | CF5S4 | 2.864 |
| OM6S1 | CM6S1 | 1.195 | OF6S1 | CF6S1 | 3.626 |
| OM6S2 | CM6S2 | 3.665 | OF6S2 | CF6S2 | 3.854 |
| OM6S3 | CM6S3 | 3.103 | OF6S3 | CF6S3 | 3.595 |
| OM6S4 | CM6S4 | 2.216 | OF6S4 | CF6S4 | 3.511 |
| OM7S1 | CM7S1 | 2.207 | OF7S1 | CF7S1 | 3.790 |
| OM7S2 | CM7S2 | 2.412 | OF7S2 | CF7S2 | 2.467 |
| OM7S3 | CM7S3 | 3.673 | OF7S3 | CF7S3 | 2.994 |
| OM7S4 | CM7S4 | 3.515 | OF7S4 | CF7S4 | 3.421 |
| OM8S1 | CM8S1 | 2.361 | OF8S1 | CF8S1 | 2.842 |
| OM8S2 | CM8S2 | 1.870 | OF8S2 | CF8S2 | 1.762 |
| OM8S3 | CM8S3 | 2.290 | OF8S3 | CF8S3 | 2.005 |
| OM8S4 | CM8S4 | 2.386 | OF8S4 | CF8S4 | 2.171 |

## APPENDIX F: Mapped Data for Analysis of Variance.

Table F.1. Table of Mapped Data using the Compared Three Logistic Functions.

| Received Speech Files (Male) | PESQ Raw Quality Score | ITU-T Rec. P.862.1 | Morfitt III & Cotanis | Obtained Function | Received Speech Files (Female) | PESQ Raw Quality Score | ITU-T Rec. P.862.1 | Morfitt III & Cotanis | Obtained Function |
|---|---|---|---|---|---|---|---|---|---|
| AM1S1 | 3.581 | 3.664 | 3.992 | 4.687 | AF1S1 | 3.205 | 3.128 | 3.579 | 4.468 |
| AM1S2 | 3.237 | 3.176 | 3.485 | 4.409 | AF1S2 | 2.166 | 1.775 | 2.139 | 3.052 |
| AM1S3 | 2.635 | 2.306 | 2.470 | 3.495 | AF1S3 | 3.107 | 2.982 | 2.554 | 3.595 |
| AM1S4 | 3.534 | 3.601 | 3.930 | 4.657 | AF1S4 | 3.296 | 3.263 | 2.790 | 3.848 |
| AM2S1 | 2.633 | 2.304 | 2.466 | 3.491 | AF2S1 | 2.416 | 2.036 | 2.605 | 3.653 |
| AM2S2 | 2.555 | 2.203 | 2.344 | 3.337 | AF2S2 | 2.687 | 2.375 | 3.174 | 4.188 |
| AM2S3 | 2.542 | 2.187 | 2.324 | 3.311 | AF2S3 | 2.828 | 2.571 | 2.441 | 3.460 |
| AM2S4 | 3.131 | 3.018 | 3.309 | 4.290 | AF2S4 | 2.718 | 2.418 | 3.901 | 4.644 |
| AM3S1 | 2.414 | 2.034 | 2.136 | 3.048 | AF3S1 | 3.051 | 2.898 | 2.551 | 3.591 |
| AM3S2 | 3.217 | 3.146 | 3.452 | 4.388 | AF3S2 | 2.617 | 2.283 | 1.953 | 2.763 |
| AM3S3 | 3.126 | 3.010 | 3.301 | 4.284 | AF3S3 | 3.513 | 3.572 | 2.751 | 3.809 |
| AM3S4 | 3.305 | 3.276 | 3.594 | 4.477 | AF3S4 | 2.685 | 2.373 | 2.283 | 3.256 |
| AM4S1 | 2.646 | 2.321 | 2.487 | 3.517 | AF4S1 | 2.276 | 1.883 | 3.156 | 4.174 |
| AM4S2 | 3.001 | 2.824 | 3.088 | 4.119 | AF4S2 | 2.805 | 2.539 | 3.331 | 4.306 |
| AM4S3 | 3.173 | 3.080 | 3.380 | 4.339 | AF4S3 | 2.515 | 2.154 | 2.541 | 3.580 |
| AM4S4 | 2.956 | 2.757 | 3.010 | 4.053 | AF4S4 | 3.041 | 2.883 | 2.785 | 3.843 |
| AM5S1 | 2.666 | 2.347 | 2.520 | 3.555 | AF5S1 | 3.144 | 3.037 | 2.605 | 3.653 |
| AM5S2 | 2.156 | 1.766 | 1.811 | 2.523 | AF5S2 | 2.679 | 2.365 | 3.508 | 4.424 |
| AM5S3 | 1.857 | 1.526 | 1.527 | 1.994 | AF5S3 | 2.825 | 2.567 | 2.515 | 3.549 |
| AM5S4 | 2.396 | 2.013 | 2.111 | 3.011 | AF5S4 | 2.718 | 2.418 | 4.146 | 4.754 |
| AM6S1 | 3.050 | 2.897 | 3.172 | 4.187 | AF6S1 | 3.251 | 3.196 | 3.700 | 4.538 |
| AM6S2 | 2.853 | 2.607 | 2.833 | 3.890 | AF6S2 | 2.663 | 2.343 | 3.947 | 4.666 |
| AM6S3 | 2.619 | 2.285 | 2.444 | 3.464 | AF6S3 | 3.706 | 3.825 | 2.761 | 3.819 |
| AM6S4 | 2.568 | 2.220 | 2.364 | 3.363 | AF6S4 | 3.374 | 3.376 | 1.710 | 2.342 |
| AM7S1 | 2.680 | 2.366 | 2.543 | 3.582 | AF7S1 | 3.547 | 3.618 | 2.547 | 3.587 |
| AM7S2 | 2.508 | 2.145 | 2.273 | 3.241 | AF7S2 | 2.811 | 2.547 | 4.032 | 4.705 |
| AM7S3 | 2.579 | 2.234 | 2.381 | 3.385 | AF7S3 | 2.061 | 1.682 | 3.587 | 4.473 |
| AM7S4 | 3.099 | 2.970 | 3.255 | 4.250 | AF7S4 | 2.683 | 2.370 | 3.575 | 4.465 |
| AM8S1 | 2.514 | 2.152 | 2.282 | 3.254 | AF8S1 | 3.612 | 3.705 | 2.605 | 3.653 |
| AM8S2 | 2.429 | 2.051 | 2.157 | 3.079 | AF8S2 | 3.301 | 3.270 | 3.579 | 4.468 |
| AM8S3 | 2.360 | 1.973 | 2.062 | 2.936 | AF8S3 | 3.293 | 3.258 | 2.139 | 3.052 |
| AM8S4 | 2.150 | 1.761 | 1.804 | 2.511 | AF8S4 | 2.718 | 2.418 | 2.554 | 3.595 |

## Appendix G: MATLAB Codes

### G.1 Script for Spectral Plot

```
>> f1=[ …..
];
>> f2=[ …..
];
>> m=[ …..
];
>> n1=[ …..
];
>> n2=[ …..
];
>> n3=[ …..
];
>> plot(f1,m,'b',f2,n1,'r',f2,n2,'g',f2,n3,'m')
>> xlabel('Frequency, Hz')
>> ylabel('Spectrum, dB')
>> title('Spectral Plot for Original & Received Speeches')
>> legend('OM1S1','AM1S1','BM1S1','CM1S1')
>>
```

### G.2 Script for Scatterand Regression Plots

```
>> x=[ …..
];
>> y=[ …..
];
```

```
>> scatter(x,y,'*','r')
>> lsline
>> xlabel('Subjective MOS')
>> ylabel('Mapped PESQ Score - MOS-LQO')
>> title('Network A Received Speeches')
>> plottools
>> plotregression(x,y)
Set(findal(gca,'Type'),'LineWidth',1.5)
```

### G.3 Script for Mapping Plots

```
>> x=[ …..
];
>> y1=[ …..
];
>> y2=[ …..
];
>> y3=[ …..
];
>> plot(x,y1,'r',x,y2,'b',x,y3,'k')
>> grid on
>> title('Improved and Existing Mapping Functions')
>> xlabel('Raw PESQ Score')
>> ylabel('Mapped PESQ Scores – MOS-LQO')
>> legend('Improved function','ITU-T function','Morfitt & Cotanis')
>>
```