NOTE TO USERS

This reproduction is the best copy available.

UMI[®]

Non-Intrusive Speech Quality Assessment in VoIP

submitted by

Ayman Radwan, B. Sc.

A thesis submitted to
Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of

Master of Applied Science (M. A. Sc.)

Ottawa-Carleton Institute For Electrical and Computer Engineering
Department of Systems and Computer Engineering
Faculty of Engineering
Carleton University
Ottawa, Ontario, Canada K1S 5B6
August, 2003

© Copyright
Ayman Radwan, 2003

Library and
Archives Canada

Bibliothèque et
Archives Canada

Published Heritage
Branch

Direction du
Patrimoine de l'édition

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 0-612-97395-6

Our file Notre référence

ISBN: 0-612-97395-6

The author has granted a non-exclusive license allowing the Library and Archives Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

PAGINATION ERROR.

TEXT COMPLETE.

ERREUR DE PAGINATION.

LE TEXTE EST COMPLET.

Abstract

VoIP is a promising technology that is expected to replace traditional telephone networks. Although VoIP is efficient, its speech quality is still less than what telephone users are accustomed to, due to packet loss, excessive delay, and clipping. Network administrators need to maintain certain level of quality by monitoring speech quality of live calls for actions to be taken.

Subjective methods are unsuitable for live calls assessment. Objective intrusive methods mandate a reference signal, therefore cannot monitor live calls. The only suitable method is the E-model that estimates the speech quality based on statistics collected from the network.

In this thesis, speech quality is estimated based only on the received degraded speech signals. Work is focused on packet loss and clipping in time which are significant and unique to VoIP. In case of packet loss, it is assumed that a lost packet is replaced by either silence or the previous packet. In each case, a method is provided to model the process then obtain an estimate of voice quality. In case of clipping, speech quality score is estimated based on the estimate of the threshold that caused the clipping. Simulation results show that speech quality can be estimated with an error less than 0.5 in MOS score in the majority of speech files examined.

Acknowledgements

I would like to thank Dr. Samy El-Hennawey and Dr. Rafik Goubran, for their continuous support and motivation during their supervision of this research. They were of great help both technically and spiritually. Special thanks to Dr. El-Hennawey for his ongoing encouragement and efforts to help me through tough times. Special appreciation for Dr. Goubran's useful guidance and advices. Due to space limitations, I will not be able to express the extent of my gratitude for both of them.

I would also like to thank Nortel networks, CITO, NSERC and Carleton University for their financial support. I would like to express my appreciation for the efforts of all my professors at Ain-shams University, Carleton University and University of Ottawa.

I would like to express my gratitude to my family back home, my aunt and her family in Canada, and all my friends in Egypt and Canada, for their encouragement during the years of this research.

At the end, I dedicate this thesis to my parents in Egypt, who were of great spiritual support for the whole years of this research. Although they are away, I can hear their prayers and words of encouragement.

Thank you ALL.

Table of Contents

ABSTRACT.....	III
ACKNOWLEDGEMENTS.....	IV
1. INTRODUCTION	1
1.1 Thesis Motivation.....	1
1.2 Problem Definition.....	5
1.3 Thesis Approach	6
1.4 Thesis Contributions.....	9
1.5 Thesis Outline.....	11
2. LITERATURE REVIEW.....	13
2.1 Voice over IP (VoIP).....	13
2.2 Advantages of VoIP	15
2.3 Challenges Facing VoIP	16
2.4 Solutions for some Impairments.....	21
2.5 Speech Quality Assessment	24
2.6 Classification of Techniques for Speech Quality Assessment	25
2.6.1 Subjective Methods	26
2.6.2 Objective Methods	26
2.6.3 Intrusive Methods	26
2.6.4 Non-Intrusive Methods	27
2.7 Existing Speech Quality Assessment Methods	28
2.7.1 Mean Opinion Score (MOS)	28
2.7.2 Perceptual Analysis/Masurement System (PAMS).....	31
2.7.3 Perceptual Speech Quality Measurement (PSQM).....	33
2.7.4 Perceptual Evaluation of Speech Quality (PESQ).....	34

2.8	Non-Intrusive Methods.....	35
2.8.1	<i>The E-Model</i>	36
2.8.2	<i>Needed Non-Intrusive Methods.....</i>	37
2.9	Summary	39
3	<u>SIMULATION ENVIRONMENT.....</u>	<u>41</u>
3.1	The Tools	41
3.2	Simulation Approach.....	42
3.2.1	<i>Proposed Non-Intrusive Method.....</i>	42
3.2.2	<i>Training Process.....</i>	44
3.2.3	<i>Evaluation of Proposed Algorithms.....</i>	46
3.3	The Speech Database.....	47
4	<u>SPEECH QUALITY ESTIMATION BASED ON PACKET LOSS</u>	<u>49</u>
4.1	Packet Loss	50
4.2	Case I: Replacement of Lost Packets with Silence.....	51
4.2.1	<i>Detecting and Locating the Occurrence of Packet Loss.....</i>	54
4.2.2	<i>Location of Lost Packets.....</i>	57
4.2.3	<i>Modeling the Effect of Packet Loss.....</i>	57
4.2.4	<i>Simulations and Results</i>	63
4.3	Case II: Replacement of Lost Packets with Previous Packets.....	68
4.3.1	<i>Detecting and Locating the Occurrence of Packet Loss:</i>	69
4.3.2	<i>Location of Lost Packets:</i>	73
4.3.3	<i>Modeling the Effect of Packet Loss.....</i>	73
4.3.4	<i>Simulations and Results</i>	80
4.4	Summary	86
5	<u>CLIPPING IN TIME.....</u>	<u>88</u>
5.1	The Clipping Effect.....	88
5.2	Modeling the Clipping Effect.....	92
5.3	Step-up Threshold Approach.....	97
5.4	Speech-Silence Duration Statistics Approach	102
5.5	Summary	107

Chapter One

Introduction

1.1 Thesis Motivation

Communication is essential these days. People need to communicate with others all over the world. They need to send all types of information such as data, voice and video. One of the most used types these days is speech [1]. Voice calls have been transmitted over traditional telephone networks for years [1]. These networks are originally designed for voice transmission, therefore they provide high quality speech conversation. Despite the high quality these networks provide, they have a major drawback. This drawback is the high cost of these calls especially for long distance calls. This high cost results from dedicating an end-to-end connection for each call. This drawback led to the spreading of alternatives such as VoIP (Voice over the Internet Protocol) [2], [3], [4]. VoIP is transmitting voice calls over packet-switched networks which use the Internet

6. CONCLUSIONS AND RECOMMENDATIONS	109
6.1 Conclusion	109
6.2 Recommendations for Future Work	111
REFERENCES.....	114

List of Tables

Table 2-1 MOS score	29
Table 2-2 PAMS score	32
Table 5-1 Threshold values versus PESQ score	94

List of Figures

Figure 3-1 Approach of non-intrusive method	43
Figure 3-2 Approach of proposed method	43
Figure 3-3 Approach of intrusive method.....	45
Figure 3-4 Approach of the evaluation procedure	47
Figure 4-1 Degraded signal (Lost packets replaced with silence packets)	53
Figure 4-2 Steps to detect the occurrence of packet loss (PLC: Replacement with silence)	56
Figure 4-3 Percentage of estimated lost energy to total energy versus degradation and the curve fitted to these points (Single packet lost)	59
Figure 4-4 Percentage of estimated lost energy to total energy versus degradation and the curve fitted to these points (Multiple packets lost)	61
Figure 4-5 Comparison between estimated quality score and PAMS score	65
Figure 4-6 Error distribution using estimated lost energy	66
Figure 4-7 Steps to detect the occurrence of packet loss (PLC: Replacement with previous packets)	72
Figure 4-8 Packet loss percentage versus PESQ score and the curve fitted to these points	75
Figure 4-9 Packet loss percentage versus PESQ score unvoiced only, voiced only, and mixed.....	77
Figure 4-10 Curves fitted for packet loss percentage versus PESQ score for unvoiced only, voiced only, and without locating the lost packets	79
Figure 4-11 Error distribution using voiced/unvoiced classification.....	83
Figure 4-12 Error distribution without using voiced/unvoiced classification	84
Figure 5-1 Effect of clipping in time due to VAD and NLP in echo cancellers.....	91
Figure 5-2 Threshold value used in VAD versus PESQ score	93
Figure 5-3 Error distribution when using the VAD threshold to estimate speech quality	96
Figure 5-4-a The received (degraded) signal	98
Figure 5-4-b Step-up threshold approach (Threshold = 0.4)	99
Figure 5-4-c Step-up threshold approach (Threshold = 0.6)	100
Figure 5-4-d Step-up threshold approach (Threshold = 0.8)	101

Figure 5-5 Degraded signals resulting from applying VAD with different threshold values to the same signal with silence-speech classification	103
Figure 5-6 Distribution of speech periods for different thresholds.....	105
Figure 5-7 Distribution of silence periods for different thresholds	106

List of Symbols

R	Transmission rating factor
R_0	The basic signal-to-noise ratio
I_s	The effect of simultaneous impairments
I_d	The impairments caused by delay
I_e	Impairments caused by low bit rates codecs
A	Expectation factor
$f(r)$	Estimated degradation
r	Percentage of total lost energy to the total energy of the signal
$S(p)$	Estimated speech quality score
p	Packet loss percentage

Acronyms

VoIP	Voice over Internet Protocol
IP	Internet Protocol
MOS	Mean Opinion Score
PAMS	Perceptual Analysis/Measurement System
PSQM	Perceptual Speech Quality Measurement
PESQ	Perceptual Evaluation of Speech Quality
DSL	Digital Speech Level Analyzer
VAD	Voice Activity Detection
NLP	Nonlinear Processor
ITU-T	International Telecommunications Union – Telecommunications
SID	Silence Insertion Descriptor

QoS	Quality of Service
PLC	Packet Loss Concealment
DMOS	Degradation Mean Opinion Score
ACR	Absolute Category Rating
DCR	Degradation Category Rating
LQ	Listening Quality
LE	Listening Effort

Protocol (IP) [3], [4]. The main reason for using VoIP is its low cost compared to traditional telephone networks especially for long distance calls [2], [4]. VoIP also has another main advantage over traditional telephone networks. This advantage is availability [4] and the efficient utilization of bandwidth. Since no connection is dedicated to one call, more calls can be up at the same time compared to traditional networks. Also, one network can carry more than one type of information together, thus only one network management is needed [2]. The main problem with VoIP is that these networks are originally designed for data transmission. These networks depend on best effort to deliver packets [3]. This technique is very efficient in data transmission but not speech transmission or real-time applications [1], [5]. Therefore, transmitting speech on packet switched networks does not guarantee the same quality offered by traditional telephone networks which users are accustomed to [2], [3], [5].

The above discussion led VoIP service providers to implement a variety of techniques to enhance speech quality offered in VoIP. This resulted in a large number of providers offering services with competing prices and different levels of quality. The problem now changed to how to assess the quality of speech offered [6]. Providers need methods to assess the performance of their services and compare it to services offered by competing providers. Engineers also need these methods to evaluate the newly developed techniques and compare it to old ones [2], [5]. Network administrators need methods to monitor the quality of speech transmitted through the networks, so they can take actions whenever the

quality of speech degrades [7]. Finally, the users also need these methods to compare the quality offered by different service providers [2].

One of the most important issues in VoIP these days is to measure the speech quality. Efforts have been focused to develop methods to measure the speech quality especially for VoIP. To measure the speech quality correctly, these methods have to reflect the human perception of speech quality. The most reliable approaches are the subjective methods. In these methods, a number of subjects (humans) rate the speech signals [2]. The average of their ratings is calculated and is considered as the quality rate of the signal. The most known subjective test is MOS (Mean Opinion Score) [6]. Details about MOS rating are provided later in the following chapter. Developers try to develop objective methods which give the same quality scores as MOS does. Objective methods are those carried out by machines, without human interference. These methods are necessary for monitoring network performance since subjective methods cannot be used for this purpose. Most of the available objective methods are intrusive in nature. In these intrusive techniques, a reference signal is injected in the network from some point and received at another point. Since the original signal is known, the received degraded signal can be rated by comparing it to the original one [5]. These techniques give relatively good estimates of MOS scores. The most reliable and widely used of these methods are PAMS (Perceptual Analysis/Measurement System), PSQM (Perceptual Speech Quality Measurement) and PESQ (Perceptual Evaluation of Speech Quality) [8]. Again,

highlights of these methods are presented in the following chapter. Another approach is called non-intrusive. In this approach, no reference signal is injected in the network. Instead, the algorithm operates on signals that are present in the network [6] or on the statistics collected from the network. The challenge in this technique is that the original signal is not known to the algorithm, so it cannot compare between original and degraded signals to assess the quality of the received signal. Some attempts are made in this approach but no robust algorithm is found. The E-model is the only method available that is considered non-intrusive [9]. This model uses the statistics collected from the network during operation, such as the packet loss rate, delay, jitter estimates and signal to noise ratio, to estimate a quality score of the signal. This method is based on statistical nature, which does not guarantee accuracy. For example, statistically if packet loss rate is high, the quality score is low. But this is not always true, since the quality of the signal is still high, if most of the lost packets are in silence periods.

Based on the above, there is a need for a method which can monitor a live call and determines the quality of this call for its whole duration. These methods are needed to alert network administrator when the quality of the call is degraded, so some action can be taken to guarantee an acceptable quality of calls all the time. This method has to assess speech quality by examining degraded signals only without any information about original signals.

1.2 Problem Definition

In the previous section, the importance of methods to measure speech quality is shown. Several methods are available now. Some of these methods are very reliable and reflect the human perception of speech quality well. Available methods are either subjective or objective. The objective methods are either intrusive or non-intrusive.

As mentioned earlier, subjective methods are the most accurate methods. MOS is the most widely used subjective method. This method reflects the human opinion so well since the quality is assessed by humans. But, this method is not applicable in live networks. Also, this method is so expensive and time consuming. Therefore this method is only used to assess the performance of speech quality when new techniques are evaluated. Objective intrusive methods are useful in network voice quality assessment in general, but they don't give individual call-by-call quality assessment. They can give a good estimate of speech quality that is offered by the network. They need a reference known signal, so they cannot monitor the quality of a real live call.

The problem is that network conditions change with time, which changes speech quality. Network administrators need to keep a certain level of quality of service offered by the network all the time. Actions, such as change of codecs, change of media, increasing the bandwidth or even dropping some calls, can be taken to maintain a certain level of quality of service offered by the network. The issue now is how to know when

any of these actions is needed. This gives rise to the need for a method which can monitor the quality of live calls. This method has to be totally objective, accurate and in the mean time does not require the injection of a reference signal. It has to determine the quality of the call using the received signal, without any information about the original sent. Summing this up, there is a need for an objective non-intrusive method to assess speech quality based on characteristics extracted from degraded signals and not statistics collected from the network. This is the main focus of the work done in this thesis.

1.3 Thesis Approach

The main purpose of this thesis is to develop an objective non-intrusive algorithm that can assess speech quality by having only access to degraded signals. Different types of impairment have different effect on speech quality. From the list of these impairments, this thesis focuses on two distinct types of impairment that are unique to VoIP. These impairments are:

- Packet loss.
- Speech clipping in time.

Each type of impairment is treated individually. In future work, it is suggested that other types of impairments be studied. Also, more than one type is modeled simultaneously, to check whether these models can be added linearly.

In this thesis, each type of impairment is applied on the signals separately. The PAMS and PESQ scores of these degraded signals are calculated using DSLA (Digital Speech Level Analyzer) by introducing both the reference and degraded signals to it. For each type of impairment, the following steps are followed using the degraded signal only:

- The detection of the occurrence of the impairment.
- The quantization of the degradation caused by this impairment.
- The mapping of the amount of degradation to PAMS or PESQ score.

PAMS and PESQ scores are not actually the ideal scores to use, as they are not perfectly designed to accommodate for the impairments examined here. MOS is a better method, but since it is not feasible here due to financial and time limitations, PAMS and PESQ scores are the best available methods to use.

Packet loss and clipping in time are the two types of impairments whose effects are modeled here. Packet loss is one of the major problems in VoIP. Packet loss usually occurs in VoIP, at peak times due to congestion. Packet loss concealment techniques are usually used in the receivers to replace lost packets with some packets to decrease its degradation effect. The packet loss concealment techniques used here are the replacement of the lost packets with constant silence packets and the replacement of the lost packets with previous packets. To model this effect, first lost packets are detected and their

locations are specified. When lost packets are replaced with silence, lost packets are detected by the sudden drop in the energy of the signal. When lost packets are replaced with previous ones, lost packets are detected by the resemblance of successive packets. The next step is to determine the degradation caused by these lost packets. In case of the first packet loss concealment technique, the degradation caused by the lost packet is determined by the amount of energy estimated in the lost packet. In case of the second technique, the amount of degradation caused by each lost packet is determined by the location of the lost packet, whether this packet falls in silence, voiced or unvoiced period. If the packet falls in silence period, it does not affect speech quality. Packets lost from voiced periods degrade the signal more than those lost from unvoiced periods. At the end, these factors are mapped to PESQ score with different weighting to estimate the quality score of degraded signals accurately.

Clipping in time usually occurs because of the usage of VAD (Voice Activity Detection) and nonlinear processor (NLP) in echo cancellers. VAD is usually used in VoIP to save bandwidth when there is silence. Echo cancellers usually estimate the echo and subtract it from the received signal. If any of these two techniques is not of very high performance, the signal is clipped in time. This means that some parts of the signal are silenced. This effect is simulated in this thesis. Depending on the threshold value of the VAD, the amount of degradation in the signal is predicted. Mapping threshold value of VAD to PESQ score, a quality score can be estimated. Two methods are suggested to

estimate the VAD threshold value. The first method depends on the procedure used in simulating the clipping effect. The second method depends on the statistics of the duration of speech and silence periods present in the degraded signal. The second method needs more investigation in future work.

1.4 Thesis Contributions

In this thesis, packet loss and clipping in time are the two types of impairment examined. Simulation results show the robustness of the proposed algorithms suggested in this thesis to accurately estimate MOS when either of the above impairment is present.

In case of replacement with silence packets, the following is achieved:

- Proposed a method to detect all lost packets which occur in speech periods. A method based on detecting packet loss by sudden drops in the energy of the signal is implemented and evaluated
 - Proposed a method to quantize the degradation caused by packet loss. The method is based on using estimated energy of the lost packet to determine the amount of degradation.
 - Formulation of a mapping to determine speech quality based on the amount of degradation obtained from the previous proposed method.
-

In case of replacement with previous packets, the following is achieved:

- Proposed a method to detect all lost packets which occur in speech periods. A method based on detecting packet loss using the resemblance of successive packets is implemented and evaluated.
- Proposed a method to quantize the degradation caused by packet loss. The method is based on using the location of lost packet to determine the amount of degradation. Packets lost during silence periods have no degradation effects. Packets lost during voiced periods have more degradation effect than those lost during unvoiced periods.
- Formulation of two mappings to determine speech quality based on the amount of degradation, one in the case when all packets are lost during voiced periods and the other when all packets are lost during unvoiced periods. When losing packets from voiced and unvoiced periods simultaneously, a weighted score between both mappings is suggested.

In both cases, lost packets during silence periods are ignored as they don't degrade the signal.

In clipping, the following is achieved:

- Suggestion of two methods to detect the clipping and quantize it. The first method is based on the method used to simulate the VAD effect. The second method is based on the statistics on the duration of speech and silence periods. More work will be done in future to formulate the second method.
- A mapping of the amount of clipping to quality score. A table mapping the estimate of the VAD threshold to quality score is deduced.

1.5 Thesis Outline

This thesis is organized as follows:

In Chapter Two, literature review about voice over IP is given. Different approaches for measuring speech quality are presented. Available methods for speech quality measurement are outlined.

In Chapter Three, simulation environment is presented. The tools and data files used in the simulations are defined. Also, the approach used in this thesis is introduced.

In Chapter Four, the first impairment examined is presented. This impairment is packet loss. Two packet loss concealment techniques are simulated and their effects are modeled. The ways to detect and locate packet loss in each case and modeling their

effects are presented. The simulations and results are also provided at the end of the chapter.

In Chapter Five, the clipping in time is examined. Simulation of this effect is presented. The methods to detect and locate this clipping and model it are shown. Two methods to quantize this clipping are suggested.

In Chapter Six, final conclusions are presented. A summary about the work done in the thesis is introduced. Points to be carried out in future work are also stated.

Chapter Two

Literature Review

In this chapter, the idea of voice over IP is introduced. Why it is believed that voice over IP will take over traditional circuit-switched telephone networks is discussed. The challenges facing VoIP is presented. Also, a summary about the available methods for speech quality assessment is introduced. Existing techniques for intrusive and non-intrusive voice quality assessment are discussed.

2.1 Voice over IP (VoIP)

VoIP is a very hot topic in telecommunications these days. It is believed that VoIP will replace the backbone of the traditional telephone networks in the near future [2]. The VoIP is the transmission of voice in packets over networks using the Internet Protocol (IP) [3]. The key to the superiority of VoIP is its low price especially in long distance

calls compared to traditional existing telephone networks [2], [4]. Actually the cost of calls in VoIP is debatable, as it can be varied depending on the quality of service provided. The advantage of it is that it can be categorized, billing high prices for high quality service and lower prices for lower quality. This cannot be done in traditional telephone networks, since making a call on traditional networks means using the whole connection for the whole duration of the call. Another factor is the availability of the network due to the high bandwidth [4] and the sharing of single connections among many calls, in contrast to traditional telephone networks where single connections are dedicated to single calls. Also in these networks data, voice and multimedia share the same network. This is an advantage for IP networks since they only require one network management. The main difference between traditional telephone networks and VoIP is the technique they use. Telephone networks are circuit switching networks, while VoIP are packet switching networks.

Circuit-switching: In this technique, a dedicated circuit connection is established between the two ends of the call. This circuit is dedicated to these two users for the whole time of the call. After the call ends, the connection is closed and the circuit is free for another call [5].

Packet-switching: In this technique, no circuit is dedicated to a certain call. Instead, the speech signals of the conversation are divided into packets. These packets are sent on

the network, and each of them is routed to the destination end through a certain route according to availability and least number of hops traversed [5].

Despite the efficiency of data networks, such as the internet, which are used by VoIP, they are not very suitable for speech transmission [1]. These networks were originally implemented for data transmission. Data transmission can tolerate long delay, retransmission of packets when discarded at some point of the network, out of order arrival of packets and varying inter-arrival delays. Speech, in contrast, is real-time application. This means, it is so sensitive to long delays. Retransmission of packets techniques cannot be used in speech applications. Also, varying delay greatly degrades the quality of speech [1]. Thus, data networks need some sophisticated techniques to ensure the quality of conversations transmitted through them. VoIP has some advantages over circuit-switching networks used by telephone companies, but also has some challenges facing it.

2.2 Advantages of VoIP

VoIP has many advantages over traditional telephone networks. The main advantage of VoIP is the cost of a call, especially the long-distance calls, which is less than that of the traditional telephone networks calls. Other advantages are:

- Availability of bandwidth and its efficient utilization [4].
-

- The extremely low bit error rate.
- The ease of expanding the networks [4].
- The format of the speech signal which is already in the form of data. Further processing of speech signals can easily be applied. Messages left on answering machines can be forwarded or saved [4].
- The easily implementation of all features of traditional telephone networks. This means other features can be integrated such as conferencing [4].
- Transmission of more than one type of information on the same network. Usually, voice, video and data are mixed on the same network. This means that they need only one network management.

2.3 Challenges Facing VoIP

As mentioned above, all these advantages are not achieved without challenges. VoIP has impairments that affect speech quality so much. The major problem of VoIP is that users are used to the good quality offered by traditional telephone networks and will always compare the speech quality offered by VoIP to that offered by traditional telephone networks [2]. Some of the most effective impairments in VoIP are [6]:

Delay: It is also called “latency”. One way latency is the time from the moment one side starts talking until this speech signal reaches the other end [1]. Delay in traditional telephone networks is usually fixed during a call. It primarily depends on the distance between the two sides [5]. It consists only of the time taken by the signal to travel through the media from one end to the other. In VoIP however, delay has more than one component. It consists of propagation delay, processing, routing and coding [5], [6]. Delay in VoIP changes according to the network usages and the time of the day [5]. Processing and coding delays are fixed, while routing depends on network conditions [10]. Long Delays can be accepted in data transmission, but not in real-time voice transmission. When delay is too long, it results in echo effects and talker overlap [1], [2], [5]. Usually, one-way delay value less than or equal to 150 ms is acceptable for professional usage. Delay values up to 400 ms are acceptable by non-professional users. They get the delay they are paying for. Delay values over 400 ms are unusable [1], [4]. These are the values of delay recommended by ITU-T [11].

Packet loss: Packet loss is a major problem in packet networks [7], [8]. Packet loss is not an issue in traditional telephone networks. Packet loss usually occurs during excessive congestion [1], [12]. When the network becomes so busy and the routers become packed with packets, routers start to discard packets [13]. It also occurs due to the excessive delay of some of the packets, that the receiver can't wait for any more and considers them lost [12]. When transmitting data, packet loss can be handled by

retransmission of lost packets. Retransmission can not be used in voice transmission [1], [14], [15] as speech is real time application and cannot stand the delay resulting from retransmission. The amount of degradation in the voice quality differs according to the distribution of packet loss [16]. For example, for low packet loss rate, a burst distribution gives better voice quality than a normal distribution, while for high packet loss rate the inverse is true [7].

Jitter: One of the problems with packet-switched networks is jitter effect [1], [13], [17]. Jitter effect is only related to packet-switched networks and is not present in circuit-switched networks. Usually, packets destined to the same end user, pass through different routes taking different amounts of time to reach their destination [18], [19]. This results in packets arriving to destination end with varying inter-arrival delay. This is the jitter effect [1], [20]. For speech signal to be understandable, it has to be played smoothly with a constant rate. Thus, with varying inter-arrival delays, this cannot be achieved. This problem is usually solved by jitter-buffers [6], [10], [18], [19], where a de-jittering algorithm is used.

Echo: Echo is the user hearing the repetition of his/her own voice [5], [14]. There are two types of echo [10]. The first type results from the four-wire to two-wire hybrid conversion and is called the electrical echo. The second type results from the sound coming out of the speakers at the destination point is picked up by the microphone and

sent back to the original end [10]. This is acoustic echo [5]. Another form of echo is the talker hearing his voice in the earpiece of the telephone. This is called side-tone and it is under control in the telephone set. This type is desirable because when the user cannot hear his/her own voice, the user gets the feeling that the other end cannot hear his/her voice too [5]. Echo is present in both kinds of circuits, but it has more effect on the conversation in the case of packet-switched networks because of the larger delay [14].

Codecs effect: To use the packet-switched networks to transmit voice, efficient compression of signals is required. This is achieved using the codecs (Coder/Decoder) [1]. Codecs are used to compress and decompress voice signals. These codecs are able to achieve a great level of compression due to the redundancy in speech signals [5]. When decoding the encoded signal, the codec tries to make the decoded signal to sound as close to the original signal as possible. This introduces some degradation that is not present in circuit-switched networks. Actually, this is a compromise between quality and bandwidth usage [14]. These days, the codecs developers are more concerned about implementing low-bandwidth codecs than better quality codecs. Though as the availability of bandwidth increases the developers should be more concerned about quality than saving bandwidth [1]. Some of the most famous ITU codecs [10], [21] are G.711 (PCM: 64kb/s) [22], G.723.1 (CELP: 5.3 & 6.3 kb/s) [23], G.729 (CELP: 8 kb/s) [24] and G.729 A (CELP: 8 kb/s) [25].

Clipping: Usually in VoIP, developers are concerned with saving bandwidth. This is why VAD (Voice Activity Detection) is usually used by codecs in the transmitting end [14], [26]. VAD usually scans the speech signal to be transmitted for the beginning and end of talk bursts [26]. When there is silence period, the VAD detects it and the transmitter does not send packets continuously as usual. Instead, silence insertion descriptor (SID) frames are transmitted with very low frequency as long as there is silence. These frames are sent to update the receiver with the characteristics of the background noise present [14]. Since packets are not sent as frequently as when there is speech, some bandwidth is saved during silence periods when there is no important information to be sent. When the silence period ends and a speech period starts, the VAD detects this again, and the transmitter starts sending speech packets by the normal rate. In the receiving end, as long as there are speech packets arriving, they are played normally. When the receiver starts to receive silence frames, it produces comfort noise according to the characteristics of the background noise it receives [14], and plays this noise until speech packets start to arrive again.

This technique is very useful in saving bandwidth, especially that silence periods usually form a large percentage of any normal conversation. But, if these techniques are not very accurate, they result in some degradation in the signal. This degradation is clipping of the speech signal [6].

There are two types of clipping. Clipping in time and clipping in amplitude. The type that results from VAD usage is clipping in time, where some speech packets are classified as silence and are not sent [6]. This results in the receiving end playing comfort noise instead of playing the speech packets that have to be played.

This shows that data networks are efficient but are not exactly designed for carrying conversation. To overcome this problem several techniques and algorithms are implemented to enhance the QoS of data networks with respect to voice transmission.

2.4 Solutions for some Impairments

Because of the above impairments, some algorithms were introduced to data networks to enhance their performance in voice transmission [4]:

Packet loss concealment techniques (PLC): As mentioned in Section 2.3, due to the nature of speech conversations, that it is a real-time application, it cannot tolerate retransmission of lost packets [1], [14], [15]. Thus packet loss is unavoidable in speech conversations over data networks [10]. Retransmission of lost packets is replaced with Packet Loss Concealment (PLC) techniques [10]. These techniques try to replace the lost packet with another packet. The simplest is to replace the lost packets with a constant packet, usually a silence packet. This technique is very simple in implementation but very inefficient. The second technique is to repeat the last received packet [4]. This technique

is also so simple, as it doesn't need any computation. It only needs the storage of the last received packet. The problem with this technique is burst packet loss. In this case, long tones will result [4]. More complicated and more efficient techniques are also available. One of them uses an interpolation between the neighboring packets. The most advanced one is using a modeled packet. This technique uses the stream of received packets to make a model for the signal, from which it can estimate the lost packet [12]. For example G.723.1 uses a packet loss concealment technique which estimates a lost packet by simulating the vocal characteristics of the last received packet and then damp the signal slowly [10], [23]. In these techniques, there is also a compromise between complexity and efficiency. The last two techniques estimate the lost packets better, but add more complexity and delay.

Anti-jitter buffer: To avoid the jitter effect, a buffer is used at the receiving end. This buffer is called anti-jitter buffer [6]. This buffer saves the received packets, until there are enough received packets to play the stream synchronously [4], [18]. This buffer is usually of variable length, so that its length can be changed according to the conditions of the network [19]. When the network is busier the length of the buffer is increased, and vice versa [4]. This buffer solves the jitter effect problem, but increases the delay [1], [7], [10].

Echo removal: Two ways to prevent echo are available. The easier is echo suppression, in which the receive path is disabled when the send path is active, i.e. the user on this end is speaking. The problem with this technique is how fast the echo suppressor can determine that the speaker has finished talking [5]. The more advanced way is echo cancellation. In this technique, the canceller remembers the sent signal and subtracts it from the received one [5], [10]. This technique works fine when the delay is short. It can be adapted to the conditions of the network, so it can remember the signal for a longer time when the network is busier, i.e. delay is longer [5]. There is a problem with this technique. Because of the finite word length and the additive transmission noise, the canceller filter will not fully model the echo channel. Therefore, there will be residual echo. This residual echo is still audible and in most cases causes annoyance to the user. Upon detecting that there is no speech from the other end (no double-talk), the canceller activates the NLP by placing a large attenuation to eliminate the residual echo. In cases, however, the double-talk detector falsely detects single-talk while there is double-talk present. In this case, clipping occurs and voice quality gets degraded.

Many other algorithms and techniques are applied to data networks to improve its performance. A lot of QoS techniques are used, such as avoiding and managing network congestion, shaping network traffic and setting traffic priorities across the network [1], [13].

2.5 Speech Quality Assessment

Because of the large number of service providers, the variety in the price and the quality of the service offered by each of them, the users need to have some ways to determine the best service offered with respect to the price they are paying. Also, developers need ways to determine the performance of their networks in voice transmission and compare it to the performance of their competitors' networks. Also networks administrators need ways to monitor the quality of their networks all the time. They need to measure the quality of live calls. If the quality of these calls degrade much, they know that there are some problems and some actions have to be done [7]. These actions can be something like changing the codecs used, changing routes or dropping some of the calls as dropping calls is much better than providing the user with bad service. These reasons led to the importance of finding methods to assess the quality of speech conversations over communication networks [27]. Speech quality assessment is vital, but what is the meaning of speech quality.

Speech quality is how the users perceive the quality of the conversation and how comfortable they are during this conversation [2]. The regular ways of measuring the performance of networks cannot be applied to assess speech quality [28]. Usually, speech quality is assessed by human based on three factors which are delay, echo and clarity [5]. The most difficult factor to define is clarity [5]. Delay and echo are easy to define and measure.

Toll quality is defined as the speech quality offered by traditional telephone networks for long-distance calls [1], [29]. Users are used to this level of quality, so they usually compare whatever quality they get with it [5]. This is the reason that toll quality is defined. People usually expect toll quality, but they can accept less than it in return of other services, as mobility and reduced cost [5].

Accordingly, special methods have to be implemented to measure the speech quality of voice transmitted on packet networks, especially the ones using IP.

2.6 Classification of Techniques for Speech Quality Assessment

Old methods to measure the speech quality for circuit switched networks are not suitable for usage with VoIP. These methods depended on the fact that degradation in speech signals is linear and time invariant [30]. Because of the usage of many non-linear algorithms in packet networks, such as codecs, these methods are not suitable for speech quality assessment for VoIP [18], [27], [30]. Engineers have been concerned with developing techniques to precisely assess the speech quality, especially for VoIP.

Different directions were followed to achieve a precise estimate of the speech quality:

2.6.1 Subjective Methods

Subjective methods are methods where subjects (persons) are used to determine the quality [6], [30]. Subjective tests are supposed to be the best method to measure speech quality, because they rate the quality according to the human perception of speech [32]. Unfortunately, these tests are very expensive and time consuming [32]. They also cannot be used to monitor networks in real time.

2.6.2 Objective Methods

Objective methods are methods where the decisions are made with machines only, without the intervention of any human [6], [31]. Although these methods are not guaranteed to give the human perception of the speech quality, they are less expensive than subjective methods and can be implemented to monitor networks [18], [33]. Also they can be used by designers to test the effect of their changes and the quality of their new algorithms [5], [18], [30]. They are used in new algorithms development and the selection between existing equipments [33], [34].

2.6.3 Intrusive Methods

Intrusive methods are methods where a known test signal is injected into the network at one end then received at another one. Both the sent (original) signal and the received

(distorted) are compared to get the quality of the speech [27], [30]. These methods require the removal of the networks under test from normal conditions [27]. This is not a practical solution, as most of the distortions are result of network loading [27].

2.6.4 Non-Intrusive Methods

Non-Intrusive methods are those where no test signal is injected into the network. These methods work on the in-service signals to predict the speech quality [27]. Since non-intrusive methods work on the in-service signal, they are very useful to monitor the quality of the network at peak time, when it is not possible to perform intrusive methods [30]. To perform an intrusive speech quality test, the test has to have access to both the sending and the receiving ends. This is not the case all the time [30], so non-intrusive tests are required. Non-intrusive tests are very practical [27], but they are not easy to implement. They need to extract some parameters from the in-service signals. These parameters have to be insensitive to the characteristics of the talker and they have to be sensitive to the types of impairments present in the network under test [27].

Another approach which is used these days to implement non-intrusive methods is assessing the speech quality based on the statistics of the network. The E-model is an example of these methods. The E-Model determines the speech quality based on some statistics of the network [7]. Examples of the factors that can be used are packet loss rate,

signal to noise ratio [7]. These techniques are easier to implement but are less efficient than those working on the received signals.

2.7 Existing Speech Quality Assessment Methods

There have been many attempts to develop speech quality assessment methods. For subjective methods, the MOS was the most famous one. Because of the importance of objective methods, many attempts are made using many different techniques and parameters [28], [35], [36], [37], [38], [39], [40], [41], [42] to implement a good objective method that can precisely predict subjective results. Three methods have been successful and recommended by most developers. These methods are PAMS, PSQM and PESQ [8]. These methods use perceptual analysis, which is a simulation of how humans perceive the quality of speech. These three methods are intrusive objective methods since the reference speech must be accessible.

2.7.1 Mean Opinion Score (MOS)

MOS is a subjective method (Section 2.6.1). Its specifications and score rating are defined in ITU-T P.800 [44], [45]. It is considered the most reliable speech quality assessment method.

In this test, a group of people (30 – 50 persons) listens to a number of speech files, which has been degraded by different network impairments [30]. Each person rates each

file. At the end, the average of these scores is calculated, and taken as the MOS score result [30], [46].

There are two tests, MOS and DMOS. In the MOS test, the subjects only listen to the degraded signals and rate them. They don't know anything about the original one [29]. In the DMOS test, the subjects listen to the original signal then the degraded one. Subjects rate the degraded files compared to the original one [36]. MOS is also known as Absolute Category Rating (ACR) [29] and DMOS as Degradation Category Rating (DCR) [36].

The score of the MOS ranges from 1 (BAD) to 5 (EXCELLENT) [12]. Toll quality is for score of 4.0 or more [8]. Table 2-1 shows the meaning of MOS scores [2], [8], [29].

<u>MOS score</u>	<u>Speech quality</u>
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 2-1 MOS score

The advantage of MOS is that the speech quality is determined by humans, so it really reflects how human hearing is affected by the impairments introduced. Also, it combines the effect of all the impairments at the same time, since humans are rating the scores, so they take into considerations all the impairments that really affect the listening of the signals [30], [32].

Despite the advantages of MOS, it also has many drawbacks and critics. The main drawback of MOS is its very high cost [30], [45]. Because of the diversity of people opinions, a very large number of subjects should participate in rating the samples, increasing the cost of the test and also the time taken [8], [45]. Also, MOS is a very time consuming procedure and needs expensive equipments which are also disadvantages of MOS. For these reasons, MOS cannot be used to monitor networks.

There are also other critics to MOS concerning its performance. MOS scores are rated by humans, so they are affected by many human characteristics, such as the mood. If the same sample is rated in more than one laboratory, the scores from all the laboratories are not guaranteed to be the same [29], [46]. Also, if the same sample is rated more than once in the different laboratories with the same group of subjects, still the score can vary from one test to the other [29].

Another critic to MOS is the effect of human memory. The location of the impairment during the sample affects the human perception of the quality. For example,

if the same signal is degraded twice to produce two different degraded signals. These two signals have the same amount of degradation, but one has this degradation in the beginning of the signal and the other has it near the end, the human rating of the quality of the second one will be less than the quality of the first [12]. Also, MOS is sensitive to the order in which the signals are introduced to the subjects [32].

Although MOS has some critics and drawbacks, it is still considered to be the most reliable test for determining the speech quality [2]. It is used to measure the performance of objective methods. This is done by comparing the results of the objective methods to the MOS scores.

2.7.2 Perceptual Analysis/Measurement System (PAMS)

PAMS is an intrusive (Section 2.6.3) [15], [18], [33] objective (Section 2.6.2) method. It was the first reliable method to assess end-to-end speech quality [8], [33]. It is one of many perceptual measuring methods, such as PSQM and PESQ. Perceptual methods determine the speech quality on the basis of the amount of degradation which is audible by humans [30].

PAMS gives two scores, Listening Quality (LQ) and Listening Effort (LE) [8], [14]. Listening Quality score is the score which determines the evaluation of the clarity of the speech. It is how understandable the speech is [8], [14]. Listening Effort score is how much effort the listener has to do to understand what was said to him [8], [14]. Both

scores have a scale from 1 to 5, with 5 being the best [14], [15]. A PAMS score of 4 or higher is considered toll quality [15]. The meaning of PAMS scores are explained in the Table 2-1 [8], [33].

<u>PAMS score</u>	<u>Listening quality</u>	<u>Listening effect</u>
5	Excellent	Complete relaxation, no effort required
4	Good	Attention necessary, no appreciate effort required
3	Fair	Moderate effort required
2	Poor	Considerable effort required
1	Bad	No meaning understood with any feasible effort

Table 2-2 PAMS score

PAMS works as follows. As an intrusive method, it injects a test signal in the network and receives it. It compares both the sent and the received signals to measure the amount of degradation [8], [19]. To compare both signals, first the two signals are time aligned [5]. This process eliminates any delay effects, and even jitter effects [18], [33]. Second step is level alignment. Both signals are brought to the same power level. Because PAMS does not know the amount of gain or loss in the system, PAMS brings both signals to a constant listening level [5], [33]. In the next step, the signals are transformed from the time domain to the frequency domain [5], [18]. These spectra are assigned to bins, to simulate the human auditory system [5], [18], [33]. After that, the results are compared and the differences are calculated [5]. These differences are called the error surface [18],

[19], [33]. These errors are mapped to predict the listening quality and listening effort scores.

PAMS determines the effect of degradations caused by coding distortion, time-clipping and packet loss [8]. PAMS does not take into account the effects of delay, jitter or attenuation [8].

2.7.3 Perceptual Speech Quality Measurement (PSQM)

PSQM is another intrusive objective method [15]. It is also one of the perceptual methods. It takes into considerations the effect of degradations on human perception of speech [8], [15], [45]. It was recommended by ITU in 1996 in recommendations P.861 [8], [20], [21], [34], [45], [47]. It was preferred over a lot of other techniques, such as LPC Cepstrum Distance Measure (CD), Information Index (II), Coherence Function (CHF) and Expert Pattern Recognition (EPR) [21]. It was originally designed to measure the quality of speech transmitted on circuit switched networks [8]. Therefore, it is sensitive to degradations, such as transmission errors [21]. It does not take into considerations the degradation effects of packet-switched networks, such as jitter and packet loss [8], [34], [45].

PSQM has the same steps as the ones of PAMS (Section 2.7.2) [5]. Though, PSQM has a rating score that is different from the one used in MOS and PAMS. PSQM rates the

speech signals on a scale from 0 to 6.5, where the lower the score the better the quality [5], [15], [20], [21]. Score of 2.0 or below is considered Toll Quality [15].

To take packet-switched networks impairments into considerations, some enhancements to PSQM algorithm were added. These enhancements are called PSQM+ [5].

2.7.4 Perceptual Evaluation of Speech Quality (PESQ)

PESQ is another intrusive objective speech quality assessment method. PESQ is a combination of the good points in both previous objective methods PAMS and PSQM [34]. It implemented the time alignment technique used in PAMS and the perceptual modeling technique used in PSQM [5], [8]. PESQ also has some new features. One of these new features is the averaging of distortion over time [8]. PESQ takes into consideration a number of impairments, such as coding distortion, errors, packet loss, delay, jitter and filtering [8]. This made PESQ a powerful method, better than PAMS and PSQM. This made PESQ to be recommended by ITU in February 2001 as ITU-T Recommendation P.862 [8].

The steps of PESQ are almost the same as the steps of PAMS and PSQM. They are only a little different. PESQ starts with level alignment between the original signal and the degraded one (Section 2.7.2). Both signals are then filtered and then aligned in time. They are transformed using the same auditory transform as the one used in PSQM. Two

errors are extracted from the transforms and then mapped to MOS score [34]. More detailed description of each step can be found in [34].

PESQ score can range from -0.5 to 4.5, although scores less than 1.0 are not likely to occur [48]. Scores of 3.8 or higher are considered Toll Quality [8].

2.8 Non-Intrusive Methods

Non-Intrusive speech quality assessment is considered a new field. Despite the importance of non-intrusive methods, only few attempts were done in this field. Examples of the attempts done are monitoring the network performance by the packets arriving and departing at a certain point of the network [43], [50]. Another one uses known signals at the receiving end, and by monitoring the degradations that happen to the received signal, such as delay, jitter and packet loss, and applying the same degradation to the known signal. Then the known signal is compared to the degraded one using one of the available intrusive methods and a prediction of the score is obtained. This technique is presented in [49].

One of the non-intrusive methods is the E-Model. It was accepted by most of the developers and defined by ITU-T in their recommendations [7], G.107 [9] and G.108 [51].

2.8.1 The E-Model

The E-Model is first implemented to model a network before implementation. It predicts the quality of the speech that is expected to result from implementing a network with the statistics introduced to the algorithm. Because of the need to non-intrusive methods, later the E-Model has been used as a non-intrusive method, as it does not require the injection of a test signal in the network. It calculates the quality of speech depending on the transmission parameters of the network [9]. The E-Model outputs an R factor that ranges from 0 to 100 [6], [7], where the higher the score the better the speech quality [12].

The idea of the E-Model depends on the fact that the impairments of the network are additive [2]. It starts with an R-value of 100, and subtracts the degradation effects of the network impairments [12]. The R factor is calculated using the following equation

$$R = R0 - Is - Id - Ie + A \quad (2.1)$$

R: Transmission rating factor. It is the combination of all the impairments effect of the network.

R0: The basic signal-to-noise ratio.

Is: The effect of simultaneous impairments.

Id: The impairments caused by delay.

Ie: impairments caused by low bit rates codecs.

A: Expectation factor. It makes up for the user expectation. For example, users can accept less quality in return for mobility.

The definition of the E-Model parameters and how to calculate them can be further explored in [9].

The E-Model is a non-intrusive method, but not a very accurate one. It calculates the speech quality based on the statistics of the network. It gives an estimate of the network performance in speech transmission according to the statistics of the network at certain time. It is a good estimate of the network performance but not very accurate for per call user perception.

2.8.2 Needed Non-Intrusive Methods

Networks administrators need some way to monitor the network and measure the speech quality all the time [7]. They need this algorithm to alert them when actions need to be taken. Some of the actions that may be required are changing the codecs used, changing routes, dropping some calls or changing the media used.

As stated previously, several methods to assess the speech quality exist. Most of these methods estimate speech quality with good precision, but none of them is applicable to

this situation. Each one of the previously mentioned methods has a problem that does not allow it to be used in this specific situation.

MOS cannot be used in live networks. Intrusive methods (PAMS, PSQM and PESQ) are good in voice quality assessment in general, but not in monitoring calls on a call-by-call basis. They can be used as a method to evaluate the performance of a network in general. They are not applicable to monitor live calls to perform actions if the quality of these calls drop. Also, these methods need to have access to both ends of a call, in order to inject the reference signal at one end and receive it at the other.

At the end, the E-model is not a very accurate method to use for a per-call basis assessment as it is based on the statistics of the network. The presence of some of the impairments does not guarantee the degradation of speech signals. For example, if there is a large percentage of packet loss and all of these losses occur in the silence period, this means that the speech signal is not degraded.

The need for a method to monitor the network for actions to be taken as a consequence of this quality assessment and the above drawbacks of the available methods lead to the need for an objective non-intrusive method which assesses the speech quality based on the in-service received signal. This means that this method has to determine the speech quality from the characteristics of the received signal without any information

about the original signal. Not a single method with these specifications is available and well documented till now.

2.9 Summary

VoIP provides a lot of enhancements over traditional telephone networks, but in order to be able to provide good speech quality transmission, it has to overcome some impairments. These impairments are packet-switched networks properties. They can be tolerated in data transmission but not in real-time applications, such as speech transmission.

Because of the large number of service providers and the variety in the quality of the service they provide and the price they offer, methods to measure the speech quality have to be developed. The most reliable methods are the subjective ones, but they are expensive and inappropriate to apply in real-time monitoring of networks. Therefore objective methods are needed to be developed. Objective methods have to simulate the human perception of the speech quality. Some intrusive methods were developed and recommended by developers and ITU-T. Because of some problems with the introduction of intrusive methods to networks in peak time, the need for non-intrusive methods arose. Non-intrusive speech quality assessment is a hot issue these days.

A lot of attempts are in progress now to develop a reliable non-intrusive speech quality method that assesses the speech quality based on characteristics extracted from the received signal without any knowledge about the original one. This is the main issue of this thesis. It is devoted to develop a method that can assess the quality of a degraded signal based on characteristics extracted from this degraded signal without any knowledge about the original one.

Chapter Three

Simulation Environment

In this chapter, the simulation environment is presented. The approach followed to reach the results presented in this thesis is introduced. Also the properties of the files used in the simulations and the tools used in the simulation are stated in this chapter.

3.1 The Tools

These simulations are performed using the Matlab™ software. These simulations are run on a Microsoft Windows™ operating system. The hardware used in the simulations is Intel P4 processor with speed 2.4 GHz.

The device used to calculate PAMS and PESQ scores for the evaluation of the proposed algorithm is DSLA (Digital Speech Level Analyzer) manufactured by Malden Electronics Ltd [52].

DSLAs are devices used to measure PAMS/PESQ scores. It is a black box. Inside this box, PAMS/PESQ algorithms are simulated. This box is interfaced to the computer. To calculate PAMS/PESQ scores, a software program specially designed for this device is used. In this program the paths to the original and degraded files are specified. The software reads both files and passes the signals to the hardware device to apply PAMS/PESQ algorithms on the signals, and estimate these scores based on the comparison of both signals.

3.2 Simulation Approach

This section provides the main approach followed in this research, and how simulations are carried out and evaluated.

3.2.1 Proposed Non-Intrusive Method

In non-intrusive methods, there is no original signal introduced in the network. The speech quality is obtained using the degraded signal only. In the proposed method, the approach followed is that the network impairments are introduced to the clean speech signal to obtain a degraded signal. This degraded signal only, is made available to the proposed algorithm. The algorithm will estimate the quality of this degraded signal without any information about the original signal. This is illustrated in Figure 3-1.

Figure 3-1 Approach of non-intrusive method

In the algorithm, the following 3 steps are followed:

- 1- Detection of the impairment
- 2- Evaluation of the amount of degradation caused by this impairment.
- 3- Estimation of the quality of this signal according to the amount of degradation calculated.

These steps are shown in Figure 3-2.

Figure 3-2 Approach of proposed method

As shown in Figure 3-2, for each type of impairment, first the impairment is introduced to the original signal, using a matlab program written specifically for each type of impairment. The exact method of simulating each type of impairment is shown in its discussion. After introducing the impairment to the signal, the occurrence of this impairment is detected based on the characteristics of the degraded signal only. Using a certain criteria based on the type of impairment, an amount of degradation is determined. A mapping of the amount of degradation to speech quality is formulated for each type of impairment. According to this mapping, the speech quality of the degraded signal is estimated based on the amount of degradation determined in the previous step.

In the simulations, a number of impairments are examined. Each individual type of impairment is introduced. Its effect is examined separately and modeled to speech quality score. The two types of impairment examined here are packet loss and clipping in time. In future work, more types of impairment will be examined. Also, modeling more than one type of impairment simultaneously will be tested.

3.2.2 Training Process

This process is performed to find the best mapping between the amount of degradation and the quality score. First, the type of impairment under test is applied to all the available speech files with different levels of degradation. PAMS and PESQ are

calculated for all the degraded files by introducing both the original and degraded signals to the DSLA device, as shown in Figure 3-3.

Figure 3-3 Approach of intrusive method

The degraded files are divided into two classes, training data and test data. The files used for training are usually almost half of the available files or less. First the impairment is applied to the all the files, and PAMS/PESQ score is measured using the DSLA device. To get the best mapping between the amounts of degradation calculated and the quality score, for each degradation level more than one file is produced. To obtain a curve that best fits the different values resulting from one degradation level, a single value is calculated for each degradation level which gives the best error distribution. This value is calculated by calculating the averages of the scores corresponding to this degradation level more than one time. Each time the average of all the files is calculated leaving one

out. The average of these averages is taken to be the value corresponding to this degradation level. This training procedure is seen to get good results.

The quality scores used in the training process are PAMS or PESQ scores calculated using the DSLA device. These scores do not accommodate well for the types of impairments used here. But, since MOS, which is the most reliable method available, cannot be used due to financial and time limitations, they are the best available methods. Their scores highly correlate to MOS scores. Also they are used alternatively through the thesis, to emphasize the idea that the algorithms proposed in the thesis do not depend on the methods used to map their results to quality score. The results shown here are mapped using PAMS and PESQ, but if better methods are found or made available, they can be used in place of PAMS and PESQ directly. For example, if MOS is made available, it can be used, which will result in better performance of the proposed algorithms.

3.2.3 Evaluation of Proposed Algorithms

Again, PAMS and PESQ are used for the evaluation of the proposed algorithms for the same reasons mentioned in the previous subsection and in the same way. To evaluate the performance of the new proposed algorithm, the network impairments are introduced to the test speech signals to obtain degraded signals. The degraded signals are introduced to the proposed non-intrusive method to estimate the speech quality of the degraded signals. Both the original and degraded signals are introduced to the DSLA device to

calculate the PAMS and PESQ scores by comparing the degraded signals to the original ones. The scores obtained from the proposed algorithm are compared to the PESQ or PAMS scores obtained from the DSLA. The errors between these two scores are calculated. The distribution of these errors determines the performance of the proposed algorithm. This is depicted in Figure 3-4.

Figure 3-4 Approach of the evaluation procedure

3.3 The Speech Database

The speech files used in these simulations are WAV files (Windows PCM). The specifications of these files were as follows:

- Sampling rate: 8000 Hz.
- Resolution: 16 bits.
- Recording type: Mono.

The speech files range from 3 seconds to 8 seconds in length. The speech files are collected from different sources. They are from different talkers both males and females. There are 248 original reference speech files. The talkers in half of these files are females and in the other half are males. Out of these 248 files, 212 are from Nortel database and 36 are recorded in the lab.

For each original file, a number of degraded signal files are produced for each type of impairment with a different level of degradation. For example, for packet loss, from each original signal 20 different degraded files are produced with various packet loss percentages. Details about the packet loss percentages are presented in the following chapter. For each type of impairment, there are about almost 1000 degraded speech files used.

Chapter Four

Speech Quality Estimation Based on Packet Loss

The main purpose of this chapter is to find an algorithm to estimate the MOS score based on the introduction of packet loss. Packet loss is the only type of impairment that is introduced to speech files here. Packet loss is treated in VoIP by packet loss concealment techniques. Two types of these techniques are explored, which are replacement of lost packets with constant silent packets and the replacement of lost packets with previous ones. The method to detect the occurrence of packet loss in each case is discussed. The quantization of the amount of degradation caused by the packet loss is shown. Finally, the modeling of packet loss to PAMS or PESQ results is explored. Simulations and results are shown in the end of each case.

4.1 Packet Loss

Packet loss is one of the major problems facing VoIP. The causes of packet loss are previously presented in Section 2.3. Packet loss is usually treated by retransmission of lost packets in packet networks. This technique can be accepted in data transmission which can tolerate large delays. Speech, however, is a real time application, which is intolerable to large delays, thus retransmission of lost packets cannot be applied. Packet loss is unavoidable in speech transmission over packet networks. To decrease the degradation in the speech quality due to packet loss, packet loss concealment techniques are developed (Section 2.4). These techniques try to replace the lost packets with other packets, in order to make packet loss invisible to users. In this chapter, two packet loss techniques are simulated and their effects modeled. These two techniques are:

- Replacement of lost packets with silence packets.
- Replacement of lost packets with the previous packet (last received packet).

There are more sophisticated techniques. However, they do not lie within the interest of the work in this thesis.

The first case is presented in Section 4.2 and the second one in Section 4.3. In each of the two sections, the replacement technique is explained and its simulation is presented.

In each case, the following three steps are taken:

- Detecting and locating the occurrence of packet loss.
- Evaluating the amount of degradation resulting from this packet loss.
- Mapping of this degradation to the PESQ score to obtain an estimate for the MOS score.

The simulations and results of each case will be presented in the following sections.

4.2 Case I: Replacement of Lost Packets with Silence

The technique used in this case is the replacement of the lost packets with zero packets. Actually, in reality they are replaced with silence packets. Usually, these silence packets are not exactly zeros. They are noise like packets with very low energy. This technique is not so efficient but its advantage is its simplicity, as it does not need any computational complexity. This technique is usually not used in today's codecs, but it is included here to be a step in further understanding the behavior of packet loss and its effect on the speech quality and help in modeling the effect of packet loss when other packet loss concealment techniques are used.

In simulating this effect, speech files are divided into packets. The length of the packets used is 10 ms. For each packet, a random number was generated between 0 and 1. If this number is greater than the packet loss percentage this packet is left unchanged, else this packet is replaced with zeros as shown in Figure 4-1. For each original file, different packet loss percentages are applied. Packet loss can occur randomly or in bursts. In the simulation, it is assumed that packet loss is uniformly distributed.

In the training process which is explained in Section 3.2.2, there were 11 original files. For each original file, 15 different packet loss percentages are applied in the range from 0 to 15% with increment of 1%. This results in 165 training files. In the testing, there are 34 original files. Also for each file, 15 different packet loss percentages are applied in the same range. This results in 510 test files.

Figure 4-1 Degraded signal (Lost packets replaced with silence packets)

In Figure 4-1, the effect of packet loss is shown. In the first part, the original clean signal is shown. In the second part, the degraded signal is shown. It can be seen that there are two packets which are lost and replaced with silence.

4.2.1 Detecting and Locating the Occurrence of Packet Loss

The first step is to detect the occurrence of packet loss. When packets are lost during speech, there is a sudden drop in the energy. Therefore, monitoring the change in the energy of the degraded speech signals can be used to determine the occurrence of packet loss.

The steps used to determine the occurrence of packet loss are:

- 1- The average energy of each packet is calculated. Energy is usually calculated as the average of the squares of the amplitudes of the samples. In these simulations the sum of the absolute value of the amplitudes of the signal is used instead of the square of the amplitude. This gives better accuracy due to the format numbers are represented with.
 - 2- An energy threshold of silence is determined. This threshold can be evaluated depending on the nature of the signal and the surroundings. This threshold points out the silence packets and the lost packets, which are replaced with zeros, as well. This threshold is useful also because it allows the algorithm to ignore packet loss when it occurs in silence periods. When packets are lost during silence periods, this should have no effect on the quality of speech.
-

- 3- Energy values less than this threshold are replaced with the threshold value. This step is used so that in the next step silence and lost packets will have the same inverse value. Also, this step avoids the division by zero.
- 4- The inverse of the energy is calculated. In this step, all silence and lost packets will have the same inverse values.
- 5- Peaks are monitored. If a packet is lost during a speech period and replaced with a zero packet, there will be a peak in the inverse of the energy graph. Whenever a peak occurs in the inverse of the energy, this indicates that a packet was lost and replaced with a zero packet. To determine if there is a peak, the value of this peak should be the inverse value of the energy threshold determined previously. Also, an assumption is made. This assumption is that no more than three successive packets can be lost; therefore the length of the peak is maximum three points.

These steps are shown in Figure 4-2.

**Figure 4-2 Steps to detect the occurrence of packet loss
(PLC: Replacement with silence)**

In Figure 4-2, the steps to detect and locate the occurrence of lost packets when replaced by silence are shown. In the first part, the degraded signal is shown. It can be seen that two packets are lost and replaced with silence. In the second part, the plot of the energy against time is shown. Also the energy threshold for silence used is shown. It can be seen that when the two packets are lost, there are two sudden drops in the energy plot. This results in two peaks in the plot of the inverse of the energy which are shown

in the third part of the figure. Since there are two peaks in the plot of the inverse of the energy, it can be said that there are two packets lost and replaced with silence.

4.2.2 Location of Lost Packets

The location of the lost packet determines its effect on the quality of the speech. Since in this case, the lost packets are replaced with zero packets, the energy of the lost packet determines the amount of degradation it causes. As the energy of the lost packet increases, the amount of degradation increases and vice versa. In reality, there is no access to the lost packets. Therefore, the estimate of the energy of the lost packet has to be calculated. A good estimation of this value is the average of the energies of the two packets before and after the lost one.

4.2.3 Modeling the Effect of Packet Loss

A mapping between the amount of packet loss that occurred and the PAMS score is the target of this step. To find this mapping, as a first step a single packet is lost from each file to estimate the degradation due to this packet loss. The steps for this mapping are as follows:

- 1- A single packet was lost from each file. For each file, more than one degraded file is produced with a single packet lost at different positions of the signal.
 - 2- PAMS score is calculated for each degraded file using the DSLA device.
-

- 3- The energy of the lost packet is estimated using the estimation previously explained (Section 4.2.2).
- 4- The percentage of the estimated lost energy to the total energy in the signal is calculated.

$$\text{Percentage} = \{(\text{Estimated lost Energy}) / (\text{Total Energy})\} * 100 \quad (4.1)$$

- 5- This percentage is plotted against the degradation in the PAMS score measured using DSLA.

$$\text{Degradation} = 5 - \text{PAMS score} \quad (4.2)$$

- 6- A curve is fitted to this plot, using the curve fitting built-in function in Matlab. The curve obtained has the following formula:

$$f(r) = -0.0026r^2 + 0.1396r + 0.0617 \quad (4.3)$$

where r is the percentage of the estimated lost energy to the total energy in the signal. The output of this curve is the estimated degradation in the speech quality score.

- 7- To calculate the estimated speech quality score the following equation is used:

$$\text{Estimated score} = 5 - \text{Estimated degradation} \quad (4.4)$$

In Figure 4-3, the percentage of the estimated lost energy to the total energy in the signal is plotted against the degradation in PAMS score. A curve is fitted to these points and is shown in the figure.

Figure 4-3 Percentage of estimated lost energy to total energy versus degradation and the curve fitted to these points (Single packet lost)

Next, more than one lost packet is considered. Different packet loss percentages are applied to each original file. The percentages range from 0% to 15% in increment of 1%. This range is enough as usually available networks have packet loss percentages much

less than 15%. In practice, a network with 15% packet loss is unacceptable. The steps are the same as the steps explained above for single packet loss. The energy of each lost packet is estimated as stated earlier in Section 4.2.2. Two concepts are investigated here. First the degradation from each lost packet is individually calculated using the curve obtained previously and then all the degradations are summed. The other concept is to add the estimated energies of all the lost packets together and use the percentage of this value to the total energy in the signal to calculate the degradation in the speech quality using the same curve. The second concept proves to be more accurate. This is expected, as the degradation caused by packet loss is not expected to be linear.

Figure 4-4 Percentage of estimated lost energy to total energy versus degradation and the curve fitted to these points (Multiple packets lost)

In Figure 4-4, the calculated percentage of the total estimated lost energy to the total energy in the signal is plotted against the degradation in the PAMS score measured using the DSLA device. Also the curve obtained before using only a single lost packet is plotted. Comparing the estimated degradation obtained using the curve to the actual degradation in the PAMS score shows that this curve gives a good estimation of the degradation in the quality score of the signal.

Examining the results obtained from the above test, some factors are deduced to enhance the performance of the algorithm and to decrease the resulting errors. These points are:

- 1- From losing one packet only in the whole signal file, the threshold of the silence could be enhanced. Actually here, the silence/speech determination is not the goal. The goal is to determine whether the lost packet affects the quality of speech or not. Therefore, this threshold is chosen to be 1.4. If the sum of the absolute amplitudes of the samples in a single packet of length 80 samples (10 ms, 8 KHz) is less than this threshold, this packet is considered silence packet.
 - 2- When the percentage of the estimated lost energy to the total energy of the signal exceeds a certain value, the PAMS score reaches 1 (Minimum PAMS score). This value is experimentally found to be 34.5%.
 - 3- When the percentage of the lost estimated energy increases, it is seen that the curve obtained before does not fit well to the actual PAMS results. From simulation results it is shown that it is better to fit another curve when the percentage exceeds 21.5%. Fitting another curve to the part where the percentage is between 21.5% and 34.5%. Therefore two polynomials are used in mapping the
-

percentage to degradation value. The functions used are as follows:

$$f(r) = \begin{cases} -0.0026r^2 + 0.1396r + 0.0617 & 0 < r \leq 21.5 \\ -0.0097r^2 + 0.6585r - 7.8238 & 21.5 < r < 34.5 \end{cases} \quad (4.5)$$

The final algorithm and results are presented in the next subsection

4.2.4 Simulations and Results

These simulations are executed to determine the performance of the proposed algorithm. In these simulations, 34 original files are used. From each file, 15 degraded files are produced, having packet loss percentage ranging from 0% to 15% resulting in 510 testing files. Packets are lost randomly. The degraded files were introduced to the new algorithm. The algorithm having access only to the degraded files predicts the quality score of the file as follows:

- 1- Detect and locate the occurrence of packet loss (Section 0).
 - 2- The energy of each lost packet is estimated (Section 4.2.2).
 - 3- The estimated energies of all the lost packets are summed to get the total estimated lost energy.
 - 4- The percentage of the total estimated lost energy to the total energy of the signal is calculated.
-

- 5- This percentage is used in the following formula to get the estimated degradation in the quality score:

$$f(r) = \begin{cases} 0 & r = 0 \\ -0.0026r^2 + 0.1396r + 0.0617 & 0 < r \leq 21.5 \\ -0.0097r^2 + 0.6585r - 7.8238 & 21.5 < r \leq 34.5 \\ 4 & r \geq 34.5 \end{cases} \quad (4.6)$$

where r is the percentage calculated and $f(r)$ is the estimated degradation.

- 6- The estimated quality score is:

$$\text{Estimated quality score} = 5 - f(r) \quad (4.7)$$

Using these calculations, the results shown in Figure 4-5 were obtained.

Figure 4-5 Comparison between estimated quality score and PAMS score

In Figure 4-5, the final mapping between the percentage of total estimated lost energy to the total energy in the signal is plotted (The red curve). To compare the estimated speech quality to the actual PAMS scores obtained using the DSLA device, the actual PAMS scores are also shown (The blue x's). From this comparison, it can be seen, that this method can give a good estimation of speech quality compared to PAMS score. To emphasize this point more, the error distribution is then calculated.

The error for each file is calculated as follows

$$\text{Error} = \text{PAMS} - \text{Calculated score} \quad (4.8)$$

The error distribution is shown in the histogram of Figure 4-6.

Figure 4-6 Error distribution using estimated lost energy

From Figure 4-6, the following is observed:

- 72.5% of the errors have absolute value less than 0.3.

- Only 12.5% of the errors have absolute value more than 0.5.
- The mean of the absolute error value is 0.24.
- The variance of the absolute error value is 0.0786.

From these simulations and results, it can be concluded that using this algorithm all packets lost and replaced with zero packets can be detected, except those occurred in silence periods or in very low energy periods which are ignored due to their ineffectiveness on the quality of speech. The exact effective packet loss percentage can be determined. The exact position of each lost packet can also be determined. A percentage of estimated lost energy and the total energy in the signal can be calculated. This percentage is used to estimate the PAMS score of this signal. This technique can estimate the PAMS score with an absolute error of value 0.3 with 72.5% confidence.

In these simulations, it is seen that the location of the lost packets in the signal has an effect on the degradation caused by these lost packets. Usually in traditional available methods, packet loss rate is used blindly, without any information about the location of lost packets. An example of these methods can be seen in the E-Model. The proposed approach using the location of lost packets to determine the amount of degradation is seen to be more efficient. This can be seen more in Figure 4-3. In this figure, losing one packet in the whole speech file results in a wide range of degradation in PAMS score,

which ranges from 0 to almost 1.2 or more, although in all cases only one packet is lost. Since all the files are of the same length, the packet loss rate is the same for all of them. Using any of the traditional methods, which use the packet loss rate blindly, results in the same estimated MOS. Using the approach proposed here, the MOS score estimated varies according to the estimated energy in the lost packet. This gives results which correlate more to the MOS scores estimated by PAMS. Also, this can be clearly seen, if there is a large packet loss rate, but most of the lost packets fall in silence periods. In this case, the proposed method gives high estimated MOS score. Traditional methods give low scores because of the large packet loss rate. The proposed method is more accurate, since lost packets during silence periods, do not degraded the signal if MOS is correctly used. From this discussion, it is seen that locating lost packets is essential to get more accurate MOS scores.

4.3 Case II: Replacement of Lost Packets with Previous Packets

The packet loss concealment technique used in this case is the replacement of the lost packets with previous ones (the last received packet). In this technique, the receiver saves the last packet received. If a packet is lost, the receiver replaces it with the saved packet. This technique is seen to perform much better than the one used in the previous section. But still there are more efficient techniques (Section 2.4). Despite this fact, this technique is still implemented in some of the codecs, which are used at present time. This technique

is used for its simplicity. It has no computational complexity. It only needs memory that has enough space to save the last received packet. This prevents the extra delay caused by more efficient techniques to estimate the shape of the lost packets.

This packet loss concealment technique is simulated and its degradation effect is modeled. In simulating this effect, a method similar to the one explained in simulating the replacement of lost packet with silence is used. The only difference is when a packet is decided to be lost, it is replaced by the previous packet in the file. By this method, an effect similar to losing packets and their replacement with the last received packet is simulated.

In the training process previously explained in Section 3.2.2, 36 original files are used. For each original file, 25 different degraded files are produced, with 5 different packet loss percentages. These percentages range from 1% to 5% with increment of 1%. For each percentage, there were 5 different ratios between voiced lost packets and unvoiced lost packets. In testing this algorithm, 29 original files are used. For each original file, also 25 different degraded files were produced with the same properties as the ones used in the training process.

4.3.1 Detecting and Locating the Occurrence of Packet Loss:

In this case, the lost packet is replaced with a previous one. Thus a repetition of a packet occurs. To determine the occurrence of packet loss in this case, the resemblance

between successive packets is examined. To examine the degree of resemblance, cross correlation between successive packets is used. Cross correlation determines the degree of resemblance between successive packets. Cross correlation gives an array of values between -1 and 1. Each one of these values determines the resemblance between the two packets with a certain shift. Whenever this value is 1, these two packets are exactly the same with this time shift. Whenever this value is -1, these two packets are also exactly the same but with opposite signs. Whenever the value is 0 or near 0, these two packets are not related to each other by any means.

In our case, the only value that matters is when there is no shift. Depending on whether this value is near 1, 0 or -1, the degree of resemblance between the two packets is determined. If a packet is lost and replaced with a previous one, the value of cross correlation between these two packets, when there is no shift, should be very high near 1 (Ideally 1). The steps used are:

- 1- Divide the signal into speech and silence periods using the energy threshold used previously (Section 4.3.1). This step is done to ignore the occurrence of packet loss during silence. During silence periods, packets are highly correlated, thus cross correlation between successive packets in silence periods is high. This step is also executed to prevent the algorithm from considering these packets as lost packets.
-

- 2- The cross correlation between each two successive packets is calculated. The matlab built-in function for cross correlation is used.
- 3- The value of cross correlation with no shift is monitored for each two successive packets.
- 4- A threshold for this value is determined. This threshold value has to be high enough, that it cannot result from two successive packets in normal speech. A value near 1 is chosen (0.8 was used in the simulation). This value is seen to be a good threshold. Many threshold values are examined during simulations. Simulation results show that 0.8 is the best value to be used. Successive packets in normal speech examined, never have cross correlation with value more than 0.8 in speech periods. This value can be exceeded during silence periods, due to the high correlation between silence packets. This is the reason for performing step 1 in the beginning of the algorithm.
- 5- Whenever the correlation is more than this threshold value, there is a lost packet replaced with the previous one, on condition that this occurred in speech period.

These steps are shown in Figure 4-7.

**Figure 4-7 Steps to detect the occurrence of packet loss
(PLC: Replacement with previous packets)**

In Figure 4-7, the steps to detect and locate lost packets when replaced with previous packets are shown. In the first part, the degraded signal is shown. In the second part, the cross correlation with no shift is shown. Also the classification of the signal into speech and silence is shown, along with the threshold (0.8) which determines the occurrence of packet loss. It can be seen that this cross correlation exceeds the threshold twice during speech periods, which indicates that two packets are lost. It can also be seen that this correlation exceeds the threshold many times during silence periods, but since the signal

is classified into speech and silence periods in the beginning of the method, these points are ignored.

4.3.2 Location of Lost Packets:

As mentioned in the previous case, the location of the lost packet determines the amount of degradation it causes to the speech quality of the signal. Here we use a different approach than the one using the estimated energy of the lost packet (Section 4.2.2). The speech signal is divided into three categories. These three categories are silence, voiced and unvoiced periods. The amount of degradation caused by a lost packet is determined by the category of the speech in which it is located.

There are many algorithms used to classify the speech signals into silence/voiced/unvoiced segments [53], [54], [55]. Any of these techniques can be used in the proposed algorithm.

4.3.3 Modeling the Effect of Packet Loss

A mapping between packet loss percentage and the PESQ score is the target of this step. To find this mapping, the following steps are followed:

- 1- Packets are lost from clean speech files with different packet loss percentages.

There are 36 original files. For each file, a number of degraded files are produced with different packet loss percentages ranging from 0% to 5%. For each original

file, there were more than one degraded file with the same packet loss percentage, but packets were lost from different locations each time.

- 2- The PESQ score is calculated for each degraded file using the DSLA device.
- 3- Packet loss percentage is plotted against PESQ score.
- 4- A curve is fitted for this plot.

The curve obtained has the following formula:

$$S(p) = -0.0015p^5 + 0.0196p^4 - 0.1017p^3 + 0.3054p^2 - 0.7418p + 4.5 \quad (4.9)$$

where $S(p)$ is the estimated speech quality score and p is the packet loss percentage.

The plot and the curve obtained are shown in Figure 4-8.

Figure 4-8 Packet loss percentage versus PESQ score and the curve fitted to these points

From Figure 4-8, it is apparent that the same packet loss percentage results in different PESQ scores with a wide range. This plot is obtained without any information about the location of lost packets. In the next step, information about the location of each lost packet was used in determining its effect on the speech quality of the signal. Whether each lost packet falls in a voiced period or unvoiced period is determined, this information is used to determine the speech quality of the signal. Packet loss percentage

was again plotted against PESQ score. In order to find the effect of the locations of the lost packets on the voice quality, the following cases are studied:

- 1- Packets are lost only during unvoiced speech.
- 2- Packets are lost only during voiced speech.
- 3- Packets are lost equally in voiced and unvoiced speech.
- 4- Packets are lost in mixed voiced and unvoiced speech with the former dominating.
- 5- Packets are lost in mixed voiced and unvoiced speech with the latter dominating.

These cases are shown in Figure 4-9.

Figure 4-9 Packet loss percentage versus PESQ score unvoiced only, voiced only, and mixed

From Figure 4-9, it is apparent that the effect of losing packets during voiced periods is greater than the effect of losing packets during unvoiced periods. This shows that it is important to determine the location of lost packets before estimating the degradation resulting from this packet loss. From these plots, two curves were obtained. One curve is obtained when all lost packets fall in unvoiced periods, and the other one is obtained when all lost packets fall in voiced periods.

The curve when all lost packets fall in unvoiced periods has the following formula:

$$S(p) = -0.0007p^5 + 0.01p^4 - 0.0583p^3 + 0.175p^2 - 0.356p + 4.5 \quad (4.10)$$

The curve when all lost packets fall in voiced periods has the following formula:

$$S(p) = -0.0023p^5 + 0.0338p^4 - 0.1908p^3 + 0.5713p^2 - 1.1818p + 4.5 \quad (4.11)$$

These two curves along with the curve obtained without using the information about the location of lost packets are shown in Figure 4-10.

Figure 4-10 Curves fitted for packet loss percentage versus PESQ score for unvoiced only, voiced only, and without locating the lost packets

From the above analysis, it is concluded that using the information about the location of lost packets in determining the estimated speech quality is essential to achieve an accurate measure. Therefore, after calculating the packet loss percentage, two scores are calculated using the two curves obtained (Unvoiced only and voiced only). The ratio of the number of packets lost during unvoiced periods and the number of packets lost during

voiced is calculated. The two scores are added with the same ratio to estimate the final speech quality of the signal.

The final algorithm and the results are presented in the next subsection.

4.3.4 Simulations and Results

These simulations are carried out to determine the performance of the proposed algorithm. In these simulations, 29 original files are used. From each file, 25 degraded files were produced, having packet loss percentage ranging from 0% to 5% with different ratios between the number of packets lost during unvoiced periods and the number of packets lost during voiced periods. Packets are lost randomly. The degraded files are introduced to the new algorithm. The algorithm, having access only to the degraded files, estimates the quality score of the file as follows:

- 1- Detect the occurrence of packet loss, using the method previously explained (Section 4.3.1).
 - 2- Determine the location of each lost packet. Determine whether this lost packet falls in the voiced or unvoiced period. The lost packet will only fall in either voiced or unvoiced periods, as the method used to detect the occurrence of packet loss ignores the silence period in the signal (Section 4.3.1). The information about
-

the voiced and unvoiced periods of each file is manually introduced to the algorithm.

- 3- The total packet loss percentage is calculated. This percentage accumulates all detected lost packets, which occurred during voiced or unvoiced periods.
- 4- Two scores are calculated using the calculated percentage in step 3 in the two curves, the curve when all lost packets fall in voiced periods and the curve when all lost packets fall in unvoiced periods. The calculation of the estimated quality score is illustrated numerically in Example 4.1.
- 5- According to the ratio between the number of packets lost during voiced periods and the number of packets lost during unvoiced periods, each of these two scores is multiplied with the proper factor.

More ways of combining the scores from the two curves will be explored in future work.

- 6- The error between the calculated score and the PESQ score for each file is calculated.

$$\text{Error} = \text{PESQ} - \text{Calculated score} \quad (4.12)$$

Example 4.1:

A speech signal consists of 400 packets. 20 packets were lost. 10 packets were lost during voiced periods, 5 were lost during unvoiced periods and 5 were lost during silence periods. Using step1, 15 lost packets are detected. The 5 packets lost during silence are ignored. Using step2, 10 lost packets are located in voiced periods and 5 in unvoiced periods. Using step3, the percentage is calculated

$$\text{Total packet loss percentage} = (15/400) * 100 = 3.75\%$$

This packet loss percentage is used in equation 4.10 and 4.11 to get two estimated quality scores S1 and S2 respectively. S1 is the quality score if all lost packets fall in unvoiced periods and S2 if all lost packets fall in voiced periods.

$$\text{The ratio of lost packets during unvoiced} = 5/15 = 1/3$$

$$\text{The ratio of lost packets during voiced} = 10/15 = 2/3$$

$$\text{The final estimated score} = 1/3 * S1 + 2/3 * S2.$$

The error distribution obtained from this simulation is shown in the histogram of Figure 4-11.

Figure 4-11 Error distribution using voiced/unvoiced classification

From Figure 4-11, the following points can be observed.

- 76.5% of the absolute value of error is less than 0.3
- Only 7.5% of the absolute value is more than 0.5

To emphasize the importance and efficiency of determining the location of lost packets before deciding its effect, another test is performed and the results from both tests are compared.

In the second test, the same steps are carried out except that the total packet loss percentage is used directly to estimate the speech quality of the signal without the information about the location of lost packets. The curve mapping the packet loss percentage to PESQ score, without locating the lost packets, is used. This is the way existing methods usually estimates MOS scores based on packet loss rates. The error distribution obtained from this test is shown in the Figure 4-12.

Figure 4-12 Error distribution without using voiced/unvoiced classification

From Figure 4-12, the following statistics can be observed.

- Only 47% of the absolute value of error is less than 0.3
- 25% of the absolute value is more than 0.5

The comparison between the results from both methods shows the importance of determining the location of lost packets before deciding its degradation effect. This algorithm gives better results than blindly deciding the degradation effect of packet loss depending only on the packet loss percentage without any information about the location of the lost packets.

As mentioned in Section 4.2.4, traditional existing methods usually assess the quality scores based on the packet loss rate without any information about the location of lost packets. The proposed algorithm is shown to give better results. This can be depicted from Figure 4-9. In this figure, for the same packet loss percentage a wide range of PESQ scores are calculated depending on the locations of the lost packets. Also from the previous comparison, it is seen that taking the effect of the location of the lost packets into account in estimating the quality score, gives better results than estimating these scores based on packet loss rates only. This shows the efficiency of the proposed algorithm compared to existing methods.

Another point in the favor of the proposed method is ignoring packet loss when occurring in silence. In the proposed algorithm, the signal is first divided into speech and silence periods. The algorithm only detects and locates lost packets which occur during speech periods, ignoring completely the silence period. This point is a good point in this algorithm, as losing packets during silence should not affect the speech quality of the signal, especially when replaced with previous ones. This is not available in existing methods.

4.4 Summary

In this chapter, the estimation of quality score based on packet loss is discussed. Two packet loss concealment techniques are simulated. In both cases, a method to detect the lost packets is suggested and simulated. All lost packets can be detected except those occurred during silence periods which are ignored deliberately. In both cases, the amount of packet loss is mapped to speech quality score. In the first case, where lost packets are replaced with zero packets, the speech quality score can be estimated with an error less than 0.3 with 72.5% confidence. In the second case where lost packets are replaced with previous packets, the speech quality score can be estimated with an error less than 0.3 with 76.5% confidence. In both cases, the probability of getting an error more than 0.5 is 12.5% and 7.5% respectively. From the discussion in the chapter, it is shown that the proposed algorithms in both cases are more efficient than available methods which

estimates MOS scores based on packet loss percentage blindly, without any information about the location of the lost packets.

Chapter Five

Clipping in Time

In this chapter, the effect of signal clipping in time is discussed. The way the clipping in time was simulated is presented. The detection of the presence of clipping is discussed. The modeling of this clipping and suggested ways to quantize this clipping are also presented.

Clipping cannot be assessed without processing the media payload. Therefore, the traditional E-model together with RTCP will not be able to take this factor in online voice quality evaluation. The work done here is focused to achieve this.

5.1 The Clipping Effect

The clipping in the signal usually occurs in two different ways. One of them is clipping in amplitude and the other one is clipping in time. The second kind of clipping is

the one which is discussed in this chapter. This kind of clipping occurs in transmitting voice over packet switched networks due to the usage of VAD (Voice Activity Detection) in the transmitting side [6]. Also, clipping occurs during double-talk when echo cancellers employ nonlinear processor (NLP) [56]. VAD is used to save bandwidth when the user in the transmitting side is not talking [14]. VAD usually scans the speech of the talker looking for the beginning and the end of speech periods [26]. When VAD detects the beginning of a silence period the transmitter does not send packets continuously as usual. Instead, silence insertion descriptor (SID) frames are transmitted with very low frequency as long as there is silence. These frames are sent to update the receiver with the characteristics of the background noise present [14]. Since packets are not sent as frequently as when there is speech, significant bandwidth is saved during silence periods when there is no important information to be sent. When the silence period ends and a speech period starts, the VAD detects this again, and the transmitter starts sending speech packets by the normal rate.

When VAD is not very precise, clipping occurs to the speech signals [14]. In this case, the VAD classifies some speech packets as silence packets. Therefore, some parts of the speech are lost due to VAD.

A similar clipping effect occurs when using a nonlinear processor (NLP) in echo cancellers [56]. Because of the finite word length and the additive transmission noise, the

canceller filter will not fully model the echo channel. Therefore, there will be uncanceled (residual) echo. This residual echo is still audible and in most cases causes annoyance to the user. Upon detecting that there is no speech from the other end (no double-talk), the canceller activates the NLP by placing a large attenuation to eliminate the residual echo. In cases, however, the double-talk detector falsely detects single-talk while there is double-talk present. In this case, clipping occurs and voice quality gets degraded. In the following discussion, while VAD is mentioned, the same argument applies equally to NLP.

Again the challenge faced in this thesis is to be able to assess voice quality in terms of estimating the MOS score of a degraded (clipped) speech signal when the original reference is not available. This chapter illustrates how this challenge is faced and tackled in a very successful way.

In simulating clipping effect, the original signal is divided into packets. The sum of the absolute of the amplitudes of the samples of each packet is calculated. A threshold value for this sum is determined. The sum value for each packet is compared with this threshold. If the sum value of the packet is equal to or greater than the threshold value, this packet is considered speech packet and is left unchanged. If this sum is less than the threshold value, this packet is considered silence packet and is replaced with all zero packet. This threshold is changed along a range of values from 0 to 6, to simulate

different levels of performance of VAD. Whenever this threshold is small, this results in less aggressive VAD. In this case, voice quality is maintained at the expense of bandwidth saving. Whenever this threshold increases the VAD becomes more aggressive, leading to more bandwidth saving at the expense of voice quality degradation. This effect is shown in Figure 5-1.

Figure 5-1 Effect of clipping in time due to VAD and NLP in echo cancellers

As shown in Figure 5-1, three different levels of VAD effect are shown. The first case, the signal is exactly the same as the original one, because no VAD is used. In the

second case, the signal is slightly clipped, as the VAD is a least aggressive, i.e. threshold value is very low. In the last case, an aggressive VAD is simulated, therefore large parts of the speech are clipped.

5.2 Modeling the Clipping Effect

We start first by finding quantitatively the relation between VAD threshold and the corresponding effect on MOS. Next, we try to estimate from the degraded speech what threshold VAD has used. This will provide us with an estimate of MOS degradation due to clipping. Again, the PESQ is used here to estimate the MOS score.

A mapping between the level of the threshold value used in VAD and the PESQ score is the target of the first step. VAD effect is introduced to a number of speech files. For each file, a number of degraded files are produced with different clipping effects. The threshold value is varied from 0 to 6 with 0.2 steps. This means that for each original file, 31 degraded files were produced. The PESQ score for each degraded file is calculated using the DSLA device. The threshold value used to simulate the effect of VAD is plotted against the PESQ score. This plot is shown in Figure 5-2.

Figure 5-2 Threshold value used in VAD versus PESQ score

In Figure 5-2, the plot of the threshold values against PESQ score is shown. Note that for each threshold value, different PESQ scores are obtained for different speech files. This is represented by the “x” points shown vertically in the figure, concentrating around the average value. This average is shown by the smooth curve in the figure. These points on the curve are shown in Table 5-1.

<u>Threshold value</u>	<u>PESQ score</u>
0	4.5
0.2	4.5
0.4	3.8002
0.6	3.1256
0.8	2.5901
1.0	2.1646
1.2	1.8603
1.4	1.582
1.6	1.3779
1.8	1.2254
2	1.1304
2.2	1.0671
More than 2.2	1

Table 5-1 Threshold values versus PESQ score

From Figure 5-2 and Table 5-1, it is obvious that increasing the threshold value used in VAD increases the degradation in the signal, resulting in decreasing the PESQ score. Also, the curve of the average values of the PESQ score shows that this curve is a good approximation of the PESQ score if the value of the threshold is known. To prove the efficiency of this approach, the PESQ score of each degraded file is estimated by mapping the threshold value used to produce this file to a PESQ score using the previously obtained curve. The error between the calculated PESQ score using DSLA and the estimated PESQ score for each file is calculated. The error distribution obtained is shown in Figure 5-3.

Figure 5-3 Error distribution when using the VAD threshold to estimate speech quality

From Figure 5-3, it can be deduced that if the threshold value used in VAD is correctly estimated, the PESQ score of the degraded file will be estimated with an absolute error value of less than 0.3 with confidence of 78.5%. Also the figure shows that the probability of getting an absolute error value of more than 0.5 is not more than 5.5%.

Based on the above results, it is clear that the remaining work comes down to the method of estimating the threshold value used in VAD. The rest of this chapter is devoted to the approaches suggested to estimate this threshold value. In Sections 5.3 and 5.4, two approaches are suggested. The first approach is based on the method used to simulate the VAD effect. The other approach is based on the statistics of the duration of speech and silence periods of the degraded signal.

5.3 Step-up Threshold Approach

In this first approach, a procedure similar to the one used to simulate the VAD effect is used to estimate the threshold value. This approach is based on the idea that applying the same procedure used to simulate the VAD effect on the degraded signal with a threshold less than that originally used to generate the degraded signal does not affect the signal. In other words, if a threshold, which is lower than the VAD threshold, is applied on the degraded signal, the resulting signal will be the same as the degraded signal. Once applying a threshold a little bit greater than the VAD threshold, the resulting signal will differ from the degraded signal.

Therefore, the procedure to simulate the VAD effect is applied repeatedly on the degraded signal with an increasing threshold. Starting from threshold value of zero, and increasing the threshold value by steps of 0.2, each time a new threshold is applied, the resulting signal is compared to the degraded signal. If there is no difference between the

two signals, the threshold is increased further and the same procedure is repeated. If the two signals start to differ, then this degraded signal resulted from a threshold value less than the one used in this step. More precisely, this signal resulted from a threshold value between this value and the threshold value used in the previous attempt.

The accuracy of VAD threshold estimate depends on the resolution of threshold step size. The step size used here is 0.2 and it is seen to be adequate. An example of this approach is shown in Figure 5-4.

Figure 5-4-a The received (degraded) signal

**Figure 5-4-b Step-up threshold approach
Threshold = 0.4**

Figure 5-4-c Step-up threshold approach
Threshold = 0.6

Figure 5-4-d Step-up threshold approach
Threshold = 0.8

In Figure 5-4, a signal is degraded with a VAD using threshold = 0.6. This degraded signal is shown in Figure 5-4-a. Using thresholds of values 0.2, 0.4 and 0.6, the resulting signal was the same as the degraded signal originally introduced to the algorithm. The iteration using threshold of value 0.4 is shown in Figure 5-4-b and the iteration using threshold of value 0.6 is shown in Figure 5-4-c, where it can be seen that the resulting signal in both cases is exactly the same as the degraded signal. The next step is to use threshold of value 0.8. This step is shown in Figure 5-4-d, where it is seen that the resulting signal is different from the degraded one. From the last two steps, it is seen that this degraded signal resulted from threshold value between 0.6 and 0.8. As a good approximation, this threshold value estimation is taken to be 0.6 which is the real threshold value that was used to produce this degraded signal. This example proves that this approach results in good estimation of the threshold value previously used to produce the degraded signal.

5.4 Speech-Silence Duration Statistics Approach

This second approach is based mainly on the statistics of the duration of the silence and speech periods in degraded signals. To emphasize the idea behind this approach, VAD effect is applied on the same signal more than one time with different threshold values. The degraded signals resulted from different threshold values are plotted. These signals are shown in Figure 5-5.

Figure 5-5 Degraded signals resulting from applying VAD with different threshold values to the same signal with silence-speech classification
a.) The original signal (Non- aggressive VAD, threshold = 0.2)
b.) Slightly degraded signal (Slightly aggressive VAD, threshold = 0.6)
c.) Degraded signal (Aggressive VAD, threshold = 1.0)
d.) Degraded signal (Very aggressive VAD, threshold = 2.0)
e.) Extremely degraded signal (Unacceptable VAD, threshold = 3.0)

Comparing the signals in Figure 5-5, it is observed that as the VAD threshold value increases the durations of the speech periods tend to decrease, at the same time the durations of the silence periods tend to increase. In other words, more speech periods have shorter duration and more silence periods have longer duration.

To emphasize the idea of this approach more, statistics on the duration of silence and speech periods are needed. In this step, simulation of VAD effect with different threshold values is applied on a number of original speech files (212 files) with different talkers (males and females) and different sentences. The duration of each speech or silence period is calculated. The distribution of these durations is calculated. Some of these distributions for different threshold values are shown in Figure 5-6 for speech periods and in Figure 5-7 for silence periods.

Figure 5-6 Distribution of speech periods for different thresholds
a.) Using non-aggressive VAD (Threshold = 0.2)
b.) Using slightly aggressive VAD (Threshold = 0.6)
c.) Using aggressive VAD (Threshold = 1)
d.) Using very aggressive VAD (Threshold = 2)
e.) Using unacceptable VAD (Threshold = 3)

Figure 5-7 Distribution of silence periods for different thresholds

a.) Using non-aggressive VAD (Threshold = 0.2)

b.) Using slightly aggressive VAD (Threshold = 0.6)

c.) Using aggressive VAD (Threshold = 1)

d.) Using very aggressive VAD (Threshold = 2)

e.) Using unacceptable VAD (Threshold = 3)

From Figure 5-6 and Figure 5-7, it becomes more obvious that as the threshold of VAD increases more speech periods have shorter duration and more silence periods have longer duration. Comparing the graphs in Figure 5-6, it can be deduced that the distribution of the duration of the speech periods tends to be narrower and more concentrated near the zero as the threshold of the VAD increases. Also from the graphs in Figure 5-7, it can be seen that the opposite occurs to the distribution of the duration of the silence periods.

5.5 Summary

In this chapter, the effect of time clipping caused by VAD or NLP on voice quality is investigated. The MOS score is estimated when only the received degraded (clipped) speech is accessible. A two-step procedure is suggested. The first step models the MOS with regard to the energy threshold used in VAD or NLP. The result of the first step is tabulated and becomes ready to use when the speech signal is received with clipping. In the second step, the VAD (or NLP) threshold is estimated and then the MOS score is obtained from the available table. Two approaches are suggested to estimate VAD/NLP threshold of the received signal. Simulation results show the robustness of both approaches.

The proposed method models the clipping effect based on the energy threshold used in VAD. This is done by estimating the VAD threshold and then mapping it to speech quality score. This algorithm is valid even if the VAD used does not use energy threshold to classify speech and silence periods. If the VAD uses other methods even frequency based ones, still when estimating speech quality of a certain period of the signal, the proposed method will treat the signal as if it is the result of a VAD using energy threshold. Accordingly it will estimate the energy threshold that must have been used in VAD if it uses the energy threshold. Therefore, the proposed method will estimate the effect of the VAD with energy based that would have produced the same resulting signal as the one really used. At the end the signal produced is the same as the one that would have been produced using the VAD with energy threshold equals the threshold estimated by the proposed method. Summing this up, the method still can estimate speech quality independently on the type of VAD used.

Chapter Six

Conclusions and Recommendations

6.1 Conclusion

The needs for a method to monitor quality of speech transmitted through live networks make it essential to implement an objective non-intrusive method to assess speech quality in VoIP. All available methods are not applicable to fulfill these needs. Subjective methods are accurate but not suitable for live networks. Objective intrusive methods give good estimates of MOS scores, but at the same time they can not estimate the quality of live calls as they mandate access to reference signal transmitted. Not many non-intrusive methods are available. The most widely known is the E-model. This method works on statistics collected from the network. Therefore it does not give an accurate estimate as some statistics may give false indication. For example, if the packet

loss rate is so high, but most of the lost packets occur during silence periods, the E-model estimates a bad quality while the quality is not significantly degraded. All this leads to focus attention to find a method to evaluate speech quality based only on the characteristics of the degraded signal received.

In this thesis, MOS scores are estimated based only on the degraded received signals. Packet loss and clipping impairments are unique to VoIP. Therefore, the research work of this thesis focuses attention to them in an attempt to provide a fully robust speech quality measure in VoIP. Because of the variation in the effect of different types of impairment present in VoIP, each type of impairment is individually investigated.

In case of packet loss, two concealment techniques are investigated in this thesis, which are replacement of lost packets with silence and replacement of lost packets with previous ones. In both cases, a method to detect and locate lost packets is suggested. After the detection of lost packets, the degradation caused by these lost packets is quantized and mapped to quality score. The location of each lost packet with regard to silence, voiced speech and unvoiced speech is taken into consideration when determining the amount of degradation it caused. The results obtained show that this technique is more accurate compared to blindly assessing speech quality based only on packet loss rate which is used by the traditional available techniques such as the E-model. In case of replacement of lost packets with constant silence packets, the results of simulation show

that MOS score can be estimated with an absolute error less than 0.3 with confidence of 72.5%. In case of replacement of lost packets with previous ones, results show that MOS score can be estimated with an absolute error less than 0.3 with confidence of 76.5%. In both cases, absolute error greater than 0.5 occurred with percentage less than 12.5%. Since real MOS scores cannot be used easily, PAMS and PESQ are employed in this thesis.

Clipping in time is a unique impairment to VoIP. It occurs due to the use of either VAD (used to save bandwidth) or NLP (used in echo cancellers). The effect of clipping is annoying when large speech portions are clipped. In this thesis, a mapping table is deduced to correlate between the energy threshold level (used in VAD or NLP) and the corresponding MOS scores. Two techniques are provided in this thesis through which the energy threshold of the received (degraded) speech signal is found. The mapping table is then used to estimate the corresponding MOS scores.

Simulation results provided in this thesis show the robustness of the suggested methods in assessing voice quality due to the aforementioned impairments.

6.2 Recommendations for Future Work

This thesis is a step in the implementation of a method to estimate MOS score based only on the received (degraded) signals. The degradation effects of packet loss and

clipping in time are modeled individually. These two types are unique to VoIP. In future work, it is recommended to model the effects of these two types of impairment collectively.

Other types of impairment are present and annoying in VoIP. Examples of these are noise with its different shapes, delay, jitter delay and echo. In future work, it is recommended to detect these impairments and model their effects individually. Also, it is recommended to model the additive effects of these types in addition to packet loss and clipping.

In packet loss, only two packet loss concealment techniques are tackled here. Other packet loss concealment techniques are suggested to be modeled in future work. In the methods provided in this thesis, it is shown that the effect of losing packets during voiced speech is greater than the effect of losing packets during unvoiced speech. Different function is used to model the effect of losing packets during each type of speech period. In the simulation, when packets are lost from both voiced and unvoiced at the same time, the two functions are used. The score from both functions are added with different weights depending on the ratio of voiced and unvoiced lost packets. This way of combining the two scores is shown to give good results. Other ways to combine the two scores are recommended to be considered. The results from these ways have to be compared to the results obtained here to determine the best way.

In clipping, two approaches, to estimate the energy threshold level (used in VAD or NLP), are suggested. The second approach needs to be further formulated.

At the end, all the methods proposed here were modeled using PAMS or PESQ scores due to the difficulty in using MOS scores. It is recommended to perform the exact modeling using MOS scores whenever a suitable database with MOS scores is made available.

References

- [1] P. Mehta, and S. Udani, "Voice over IP," *IEEE Potentials*, vol. 20, no. 4, pp. 36 – 40, Nov. 2001.
- [2] Jin-suk Kang, Young-do Kang, In-ho Na, Yeon-sung Choi, and Jang-hyung Kim, "A study of subjective speech quality measurement over VoIP network," in *Proc. IEEE Conf. on Info-tech and info-net*, vol. 5, no. 3, pp. 311 - 316, Nov 2001.
- [3] Li Zheng, Liren Zhang, and Dong Xu, "Characteristics of network delay and delay jitter and its effect on voice over IP (VoIP)," in *Proc. IEEE Int. Conf. on Communications*, vol. 1, pp. 122 - 126, June 2001.
- [4] Torsten Schulz, "Voice over IP," in Eicon Networks white papers, Feb. 2001.
- [5] P. Denisowski, "How does it sound?," *IEEE Spectrum*, vol. 38, no. 2, pp. 60 - 64, Feb. 2001.
- [6] Telchemy technical reference, "What is 'Voice Quality'," technical notes.

-
- [7] A. D. Clark, "Modeling the effects of burst packet loss and recency on subjective voice quality," Telchemy technical notes, 2001.
 - [8] Vennie Servis, "Measuring voice quality over VoIP networks," The Tolly group news, Dec. 2001.
 - [9] ITU-T Rec. G. 107, "The E-model, a computational model for use in transmission planning," Mar 2003.
 - [10] T. J. Kostas, M. S. Borella, I. Sidhu, G. M. Schuster, J. Grabioca, and J. Mahler, "Real-time voice over packet-switched networks," *IEEE Network*, vol. 12, no. 1, pp. 18 - 27, Feb. 1998.
 - [11] ITU-T Rec. G. 114, "One-way transmission time," May 2003.
 - [12] Luis F. Ortiz, "Solving QoS in VoIP: A formula for explosive growth," Brookout Technology Special Report, July 2001.
 - [13] B. Keepence, "Quality of Service for voice over IP," *IEE Colloquium on Service Over the Internet – What does Quality cost? (Ref. No. 1999/099)*, pp. 1 – 4, June 1999.
 - [14] MCK white papers, "MCK & Voice Quality Testing (VQT)," 2001.
 - [15] B. S. De, P. P. Joshi, V. Sahdev, and D. Callahan, "End-to-end voice over IP testing and the effect of QoS on signaling," in *Proc. the 35th symposium on System Theory*, pp. 142 – 147, March 2003.
-

-
- [16] Lingfen Sun, and E. C. Ifeachor, "Perceived speech quality prediction for voice over IP-based networks," in *Proc. IEEE Int. Conf. on Communications*, vol. 4, pp. 2568 - 2572, May 2002.
 - [17] A. Lakaniemi, J. Rosti, and V. I. Raitanen, "Subjective VoIP speech quality evaluation based on network measurement," in *Proc. IEEE Int. Conf. on Communications*, vol. 3, pp. 748 - 752, 11 - 14 June 2001.
 - [18] A. W. Rix, and M. P. Hollier, "The perceptual analysis measurement system for robust end-to-end speech quality assessment," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, pp. 1515 - 1518, June 2000.
 - [19] A. W. Rix, R. Reynolds, and M. Hollier, "Robust perceptual assessment of end-to-end audio quality," *IEEE Workshop Applications of Signal Processing to Audio and Acoustics*, pp. 39 - 42, Oct. 1999.
 - [20] Spirent Communications white paper, "Automated testing for quality of speech: Abacus provides load testing for packetized speech," 2001.
 - [21] Spirent Communications white paper, "Using PSQM to test packetized speech," Mar. 2000.
 - [22] ITU-T Rec. G.711, "Pulse Code Modulation (PCM) of voice frequencies," Nov. 1988.
-

-
- [23] ITU-T Rec. G. 723.1, "Dual rate speech coder for multimedia communication transmitting at 5.3 and 6.3 Kbit/s," Mar. 1996.
 - [24] ITU-T Rec. G.729, "Coding of speech at 8 Kbit/s using Conjugate Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)," Mar. 1996.
 - [25] ITU-T Rec. G.729 Annex A, "Reduced complexity 8 Kbit/s CS-ACELP speech coder," Nov. 1996
 - [26] R. V. Prasad, A. Sangwan, H.S. Jamadagni, M. C. Chiranth, R. Sah, and V. Gaurav, "Comparison of voice activity detection algorithms for VoIP," in *Proc. Seventh International Symposium on Computers and Communications*, pp. 530 – 535, July 2002.
 - [27] P. Gray, M. P. Hollier, and R. E. Massara, "Non-intrusive speech-quality assessment using-vocal tract models," in *Proc. IEEE on Vision, Image and Signal Processing*, vol. 147, no. 6, pp. 493 - 501, Dec 2000.
 - [28] N. B. Seitz, S. Wolf, S. Voran, and R. Bloomfield, "User-oriented measures of telecommunication quality," *IEEE Communications Magazine*, vol. 32, no. 1, pp. 56 - 66, Jan. 1994.
 - [29] An N.E.T. White Paper, "The Mean Opinion Score,".
-

-
- [30] A. W. Rix, M.P. Hollier, and P. Gray, "Predicting speech quality of telecommunications systems in a quality differentiated market," in *Proc. 6th IEE Conf. on Telecommunications*, pp. 156 - 160, April 1998.
 - [31] Jin Liang, and R. Kubichek, "Output-based objective speech quality," in *Proc. IEEE 44th Vehicular Technology Conf.*, vol. 3, pp. 1719 – 1723, June 1994.
 - [32] Shuzhen Wu, and Louis C. W. Pols, "A distance measure for objective quality evaluation of speech communication channels using also dynamic spectral features," in *Proc. Institute of Phonetic Sciences, University of Amsterdam*, 1996.
 - [33] Psytechnics white paper, "PAMS: Measuring speech quality over networks as the customers hear it," May 2001.
 - [34] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 749 – 752, May 2001.
 - [35] Yang Zhen, and Li Canwei, "A new speech quality assessment method based on human perception and its application," in *Proc. 5th Int. Conf. on Signal Processing*, vol. 2, pp. 666 – 669, Aug. 2000.
-

-
- [36] L. Thorpe, and Wonho Yang, "Performance of current perceptual objective speech quality measures," in *Proc. IEEE on Speech Coding*, pp. 144 – 146, June 1999.
- [37] K. T. Petersen, J. A. Sorensen and, S. D. Hansen, "Objective speech quality assessment of compounded digital telecommunication systems," in *Proc. IEEE First Workshop Multimedia Signal*, pp. 137 – 142, June 1997.
- [38] K. T. Petersen, S. D. Hansen, and J. A. Sorensen, "Speech quality assessment of compounded digital telecommunication systems; Perceptual dimensions," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 2, pp. 1375 – 1378, April 1997.
- [39] M. Hansen, and B. Kollmeier, "Using a quantitative psychoacoustical signal representation for objective speech quality measurement," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 2, pp. 1387 – 1390, Apr. 1997.
- [40] N. Magotra, M. Kirstein, S. Sirivara, and T. Hamill, "Quantitative and qualitative (subjective) perceptual measures for speech processing applications," *Record of the 13th Asilomar Conf. on Signals, Systems and Computers*, vol. 1, pp. 766 – 769, Nov. 1996.
-

-
- [41] M. M. Meky, and T. N. Saadawi, "A perceptually-based objective measure for speech coders using abductive network," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, pp. 479 – 482, May 1996.
 - [42] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. on Communications, Computers and Signal Processing*, vol. 1, pp. 125 – 128, May 1993.
 - [43] P. Skelly, and M. Li, "EIPMon: an enhanced IP network monitoring tool," *IEEE Workshop Internet Applications*, pp. 20 – 27, July 1999.
 - [44] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," 1996.
 - [45] International Engineering Consortium, "Voice Quality (VQ) in converging telephony and Internet Protocol (IP) networks," 2003.
 - [46] H. Knoche, H. G. De Meer, and D. Kirsh, "Utility curves: Mean opinion scores considered biased," in *Proc. Seventh Int. Workshop on Quality of Service*, pp. 12 – 14, Jun. 1999.
 - [47] ITU-T Rec. P.861, "Objective quality measurement of telephone-band (300 – 3400 Hz) speech codecs," 1996.
-

-
- [48] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for an end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Feb. 2001.
 - [49] A. E. Conway, "A passive method for monitoring voice-over-IP call quality with ITU-T objective speech quality measurement methods," in *Proc. IEEE Int. Conf. on Communications*, vol. 4, pp. 2583 – 2586, May 2002.
 - [50] J. M. Matta, and A. Takeshita, "End-to-end voice over IP quality of service estimation through router queuing delay monitoring," in *Proc. IEEE Global Telecommunications Conf.*, pp. 2458 – 2462, Nov. 2002.
 - [51] ITU-T Rec. G.108, "Application of the E-model: A planning guide," Sept. 1999.
 - [52] Malden Electronics Ltd, www.malden.co.uk
 - [53] S. Knorr, "Reliable voiced/unvoiced decision," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 27, no. 3, pp. 263 – 267, Jun 1979.
 - [54] L. Rabiner, and M. Sambur, "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 25, no. 4, pp. 338 – 343, Aug 1977.
 - [55] B. Atal, and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 201 – 212, Jun 1976.
-

- [56] ITU-T Rec. G. 168, "Digital network echo cancellers," Jun. 2002.
-