UNIVERSIDADE DE TRÁS-OS-MONTES E ALTO DOURO

Methods for Quality Enhancement of Voice Communications Over Erasure Channels

Tese de Doutoramento em Informática

Filipe dos Santos Neves

Orientador: Salviano Filipe Silva Pinto Soares
Co-orientador: Pedro António Amado de Assunção

VILA REAL, 2015

Universidade de Trás-os-Montes e Alto Douro

Methods for Quality Enhancement of Voice Communications Over Erasure Channels

Tese de Doutoramento em Informática

Filipe dos Santos Neves

Orientador: Salviano Filipe Silva Pinto Soares
Co-orientador: Pedro António Amado de Assunção

Vila Real, Julho de 2015

Methods for Quality Enhancement of Voice Communications Over Erasure Channels

Tese de Doutoramento em Informática

Filipe dos Santos Neves

Orientador: Salviano Filipe Silva Pinto Soares
Co-orientador: Pedro António Amado de Assunção

Tese submetida à

UNIVERSIDADE DE TRÁS-OS-MONTES E ALTO DOURO
para satisfação dos requisitos de obtenção do grau de
DOUTOR em
Informática
DR - I série - A, Decreto-Lei n.º 74/2006 de 24 de março e
Regulamento de Ciclo de Estudos Conducente ao Grau de Doutor na UTAD
DR, 2.ª série - N.º 149 - Regulamento n.º 472/2011 de 4 de agosto de 2011
DR, 2.ª série - N.º 244 - Declaração de retificação n.º1957/2011 de 22 de dezembro de 2011

Vila Real, Julho de 2015

Scientific advisors:

Salviano Filipe Silva Pinto Soares

Professor Auxiliar do
Departamento de Engenharias da Escola de Ciências e Tecnologia
Universidade de Trás-os-Montes e Alto Douro

Pedro António Amado de Assunção

Professor Coordenador da
Escola Superior de Tecnologia e Gestão
Instituto Politécnico de Leiria

Doctoral Commitee

President

Professor Doutor Vitor Manuel de Jesus Filipe

Professor Associado com Agregação da Universidade de Trás-os-Montes e Alto Douro

Members

Professor Doutor Paulo Jorge dos Santos Gonçalves Ferreira

Professor Catedrático da Universidade de Aveiro

Professor Doutor Carlos Manuel Gregório Santos Lima

Professor Auxiliar da Universidade do Minho

Doutor Victor Manuel Letra Macedo Marques

Consultor Sénior da PT Inovação e Sistemas

Professor Doutor Manuel José Cabral dos Santos Reis

Professor Associado com Agregação da Universidade de Trás-os-Montes e Alto Douro

Professor Doutor Salviano Filipe Silva Pinto Soares

Professor Auxiliar da Universidade de Trás-os-Montes e Alto Douro

Professor Doutor Pedro António Amado Assunção

Professor Coordenador do Instituto Politécnico de Leiria

Dedicated to those who pointed me the way

my parents,
Maria and Manuel
(in memoriam).

Acknowledgments

The present work would not be accomplished without the support and contribution of many institutions and people. To all of them, I would like to express my gratitude.

First of all, I would like to thank my scientific advisor Salviano Soares and co-advisor Pedro Assunção for their constant support, fruitful discussions, exigence, patience and encouragement. To Salviano, a special thanks for the opportunity to work in the area, for accepting me to work with him, for his vision and guidance, which permitted to successfully carry out the work herein presented. To Pedro, a special thanks for his indispensable help namely on revising all the documents and for his suggestions. Without these advisors' complementarity, the work presented in this thesis would not be possible.

I also want to thank University of Trás-os-Montes e Alto Douro (UTAD) and Polytechnic Institute of Leiria, where I work as a lecturer, for the opportunity to engage in the PhD program and for all the logistic and support that I needed to carry out the work presented in this thesis. Additionally, I also would like to thank Instituto Politécnico de Leiria / Escola Superior de Tecnologia e Gestão for relieving my teaching duties, without which this research work would not have been possible in due time.

I also acknowledge the support of Fundação para a Ciência e a Tecnologia (FCT), under the program "Programa de Apoio à Formação Avançada de Docentes do Ensino Superior Politécnico" (PROTEC) (ref. SFRH/BD/49856/2009).

I want to thank Portugal Telecom Inovação (PTIn) for recognising the importance of our area of investigation, for accepting our project application to the "Plano Inovação" and partially supporting it. This collaboration resulted in a fruitful partnership that permitted the development of the Project "e-VoIP". Thanks also to PTIn for the resources used in the experimental work, which permitted to validate the voice quality evaluation model. In this context, special thanks must be given to Manuel Aguiar, Victor Marques and Filipe Tavares for supervising the project and to Simão Cardeal for providing some of the experimental data.

I also would like to acknowledge the logistic and financial support provided by Instituto de Telecomunicações (Leiria branch). Special thanks also to Lino Ferreira, Sylvain Marcelino and Pedro Correia, members of IT, for their help and friendship.

I also thank to the department coordinators Patrício Domingues and Vitor Távora for their understanding and enabling the flexibility that permitted me to combine academic activities with research work. Thanks also to Beatriz Piedade, with whom I share academic activities, since her help allowed me to have more time for writing this thesis and to Miguel Frade and Olga Craveiro for some tips and friendship.

I also would like to thank Goreti Monteiro, English lecturer at IPLeiria/ESTG, for the voice samples that she allowed me to record for use in this work.

In a personal plan, I wish to express my gratitude to my friends of the "Mambré fraternity" for their vital presence, which helped me to support the solitude. A special acknowledgment to my friends Aldina & José, Cristina Costa, Fernando Pascoal, María Ángeles, Marília Morgado and my cousin Cathy for their presence, care and encouragement.

Finally, last but not the least, a special tenderness for my beloved father (*in memorium*) for his always affable support and understanding on depriving himself of my presence.

Additionally (why not?) my gratitude to the Creator for the offered wonders, since they fed my soul, specially during the hard moments.

UTAD, Vila Real — Filipe dos Santos Neves
17 July 2015

Abstract

This thesis presents a research work carried out by the author in the context of the Quality of Experience (QoE) in error-prone voice communication systems. Relevant research problems are identified and so the motivation for the investigation herein presented is established, starting from the disturbances that contribute to the impairment of the intelligibility experienced by users. Then, a review of the most important techniques currently found in the literature to enhance voice quality in communication systems prone to transmission errors and data loss is presented. Packet Loss Concealment (PLC), Quality of Service (QoS) and packet prioritisation are addressed for this purpose. In the context of voice quality enhancement it is necessary to assess how much effective an enhancing technique is. Thus, the most significant methods used for telephony voice quality evaluation are described, considering the human subjective factors. Subjective methods of voice quality evaluation are reviewed and the relevant terminology is established. Then, objective methods, that are suitable for computational implementation to compute a score of the voice quality as it would be scored by an *average* subject are also reviewed. The most widely accepted and standard ones are studied, most of them released by the International Telecommunication Union (ITU). Special emphasis is given to Perceptual Evaluation of Speech Quality (PESQ), that uses a reference input signal and to E-Model, that essentially uses the characteristic parameters to provide an estimate of the transmission quality, taking into account the entire communication pathway of an end-to-end telephony system.

A practical model for voice quality evaluation was investigated and validated according to the ITU Telecommunication Standardisation Sector (ITU-T) Rec. P.564 requirements. The results show that such a model complies with the therein specified class 2 of accuracy. Two linear interpolation algorithms permitting to reconstruct lost samples of voice signals transmitted through erasure channels are investigated and proposed as means to enhance the voice quality. After defining the concept of dimension in the resolution of a problem, as well as the key parameters that condition such kind of problems, the maximum dimension discrete version of the Papoulis-Gerchberg algorithm and a minimum dimension algorithm are described and used to implement a method of voice signal reconstruction. The results permit to conclude that these algorithms are suitable to recover missing samples when erasures exhibit an interleaved geometry and consider the interleaving structure of the samples in the source as a strategy to put, *a priori*, the problem in a well-conditioning point by judiciously choosing the key parameters.

This thesis also describes a research study concerning voice packet classification according to the importance each one has in the overall voice quality. It aims to give them different priorities and preferentially lose those of less importance in networks with the capability of implementing channels with different priorities. A classification algorithm based on a dynamic programming approach is proposed and mathematically formulated to define a packet prioritisation scheme for transmission over priority networks. The results show that, under random packet loss, prioritised signals are less distorted and have better Mean Opinion Score (MOS) than signals sent without any priority. A novel technique combining this method and the Papoulis-Gerchberg algorithm is proposed with the aim of exploring synergies in the reconstruction of voice signals. The results show a decrease in the number of the Papoulis-Gerchberg iterations as well as a decrease in the reconstruction error. Overall, this novel technique contributes to enhance the performance of the signal reconstruction when using the maximum and minimum dimension processes, which can find useful applications in enhancing the QoE in voice communications.

Key Words: Signal reconstruction, voice quality, MOS enhancement, VoIP, QoE, packet prioritisation.

Resumo

Esta tese apresenta um trabalho de investigação levado a cabo pelo autor no contexto da Qualidade de Experiência (Quality of Experience, QoE) em sistemas de comunicação de voz sujeitos a erros. Nela se identificam os problemas de investigação mais relevantes, dos quais decorre a motivação para a investigação apresentada, a começar pelas perturbações que contribuem para a degradação da inteligibilidade experimentada pelos utilizadores. De seguida é apresentada uma revisão das técnicas de melhoria da qualidade de voz em sistemas de comunicação sujeitos a erros de transmissão e perdas de dados, presentes actualmente na literatura. Neste contexto são abordadas as técnicas de dissimulação de perdas de pacotes (Packet Loss Concealment, PLC), de qualidade de serviço (Quality of Service, QoS) e de priorização de pacotes. Como consequência, surge a necessidade de avaliar a eficiência de uma determinada técnica. Neste âmbito, são descritos os métodos mais importantes actualmente utilizados para a avaliação da qualidade de voz telefónica tendo em conta os factores humanos de avaliação, inerentemente subjectiva. São assim apresentados os métodos subjectivos de avaliação da qualidade de voz, estabelecendo-se ao mesmo tempo a terminologia mais relevante. São também apresentados os métodos objectivos adequados a uma implementação computacional capaz de calcular uma pontuação relativa à qualidade de voz tal como seria pontuada por um utilizador *médio*. Neste contexto são estudados os métodos padrão mais amplamente aceites, a maior parte dos quais disponibilizados pela União Internacional de Telecomunicações (International Telecommunications Union, ITU). É dada especial ênfase ao denominado método PESQ (Perceptual Evaluation of Speech Quality), que usa à entrada um sinal de referência, e ao E-Model, que usa essencialmente os parâmetros que caracterizam o caminho extremo a extremo de uma ligação telefónica, incluindo os respectivos componentes, para determinar

uma estimativa da qualidade de transmissão.

No trabalho aqui apresentado é investigado um modelo prático de avaliação da qualidade de voz, validado de acordo com a recomendação P.564 do sector de padronizações da ITU (ITU Telecommunications Standardization Sector, ITU-T). Os resultados obtidos mostram que esse modelo cumpre os requisitos especificados para ser incluído na classe 2 de precisão. Como forma de melhorar a qualidade de voz, são também investigados dois algoritmos de interpolação linear que permitem reconstruir amostras perdidas em sinais de voz transmitidos através de canais com apagamentos. Após ser definido o conceito de dimensão na resolução de um problema, bem como os parâmetros chave que condicionam este tipo de problemas, são descritos e usados tanto um algoritmo de dimensão mínima como um algoritmo de dimensão máxima –a versão discreta do algoritmo de Papoulis-Gerchberg, de modo a implementar um método de reconstrução do sinal de voz. Os resultados obtidos permitem concluir que estes algoritmos são adequados para recuperar amostras perdidas devido a apagamentos que exibam uma geometria entrelaçada e assim considerar uma estrutura de amostras entrelaçada na fonte, como estratégia para colocar, *a priori*, o problema num ponto de bom condicionamento mediante escolha adequada dos parâmetros chave.

Esta tese descreve ainda um estudo relativo à classificação de pacotes de voz, de acordo com a importância que cada um tem na qualidade de voz global. O estudo tenciona atribuir diferentes prioridades aos pacotes de modo a fazer perder preferencialmente os de menor importância em redes com a capacidade de implementar canais com diferentes prioridades. É assim proposto, e matematicamente formulado, um algoritmo de classificação de pacotes baseado numa abordagem de programação dinâmica de modo a definir um esquema de priorização para transmissão em redes que implementem prioridades. Os resultados mostram que perante perdas aleatórias de pacotes, os sinais sujeitos a priorização são menos distorcidos e apresentam uma melhor pontuação média de opiniões (Mean Opinion Score, MOS) que os sinais não sujeitos a qualquer priorização. É ainda proposta uma nova técnica que combina este método com o algoritmo de Papoulis-Gerchberg com o objectivo de explorar sinergias na reconstrução de sinal. Os resultados mostram uma diminuição no número de iterações deste algoritmo bem como uma diminuição no erro de reconstrução. No global, esta técnica contribui para melhorar o desempenho da reconstrução de sinal ao usar os algoritmos de dimensão máxima e de dimensão mínima, o que poderá revelar-se útil em aplicações de melhoria de QoE em comunicações de voz.

Palavras-chave: Reconstrução de sinal, qualidade de voz, melhoria MOS, VoIP, QoE, priorização de pacotes.

Table of Contents

List of Tables

List of Figures

Acronyms

AAC Advanced Audio Coding.

AAC-LD Advanced Audio Coding - Low Delay.

ACB Adaptive Codebook.

ACELP Algebraic CELP.

ACR Absolute Category Rating.

AGC Automatic Gain Control.

AMR Adaptive Multi-Rate.

AMR-NB Adaptive Multi-Rate Narrowband.

AMR-WB Adaptive Multi-Rate Wideband.

AODV Ad-hoc On-Demand Distance Vector.

ARCEP Autorité de Régulation des Communications Électroniques et des Postes.

BER Bit Error Rate.

BSD Bark Spectral Distortion.

BYOD "Bring Your Own Device".

CCR Comparison Category Rating.

CCS Common Channel Signaling.

CDMA Code Division Multiple Access.

CELP Code-Excited Linear Prediction.

CETVSQ Continuous Evaluation of Time Varying Speech Quality.

CLR Circuit Loudness Rate.

CQS Conversational Quality Subjective.

CRN Cognitive Radio Network.

DCR Degradation Category Rating.

DECT Digital Enhanced Cordless Telecommunications.

DFT Discrete Fourier Transform.

DiffServ Differentiated Services.

DP Dynamic Programming.

DSP Digital Signal Processing.

DSR Dynamic Source Routing.

DTX Discontinuous Transmission.

EDF Earliest-Deadline-First.

EFR Enhanced Full Rate.

EIA Electronic Industries Alliance.

EVRC Enhanced Variable Rate Codec.

FCT Fundação para a Ciência e a Tecnologia.

FEC Forward Error Correction.

FoIP Fax over IP.

FPGA Field-Programmable Gate Array.

GPRS General Packet Radio Service.

GSM Global System for Mobile communications.

ICB Innovative Codebook.

IDFT Inverse of the Discrete Fourier Transform.

iLBC internet Low Bitrate Codec.

IP Internet Protocol.

iSAC internet Speech Audio Codec.

ISDN Integrated Services Digital Network.

ITU International Telecommunication Union.

ITU-T ITU Telecommunication Standardisation Sector.

LAN Local Area Network.

LPC Linear Predictive Coding.

LQE Listnen-Quality Estimated.

LQO Listening-Quality Objective.

LQS Listening-Quality Subjective.

LSMR Listener Sidetone Masking Rating.

LSTR Listner Side Tone.

MDC Multiple Description Coding.

MOS Mean Opinion Score.

MP3 MPEG-2 Audio Layer III.

MPEG Moving Picture Experts Group.

MSE Mean Square Error.

NR Noise Reduction.

OCF Older-Customer-First.

OECD Organization for Economic Co-operation and Development.

OLR Overall Loudness Rate.

OLSR Optimised Link State Routing.

P.OLQA Perceptual Objective Listening Quality Assessment.

PAMS Perceptual Analysis Measurement System.

PAQM Perceptual Audio Quality Measure.

PBX Private Branch Exchange.

PCM Pulse Code Modulation.

PDC-FR Personal Digital Cellular - Full Rate.

PESQ Perceptual Evaluation of Speech Quality.

PhD Philosophiae Doctor.

PLC Packet Loss Concealment.

POTS Plain Old Telephone Service.

PSNR Peak Signal to Noise Ratio.

PSQM Perceptual Speech Quality Measure.

PSTN Public Switched Telephone Network.

PTIn Portugal Telecom Inovação.

QCELP Qualcomm CELP.

qdu quantisation distortion unit.

QoE Quality of Experience.

QoS Quality of Service.

RF Radio Frequency.

RLR Receive Loudness Rate.

RMSE Root Mean Square Error.

RNN Random Neuronal Network.

ROHC Robust Header Compression.

RS Reed-Solomon.

RTP Real-time Transport Protocol.

SLR Send Loudness Rate.

SNR Signal-to-Noise Ratio.

SNRseg Segmental Signal-to-Noise Ratio.

SOR Successive Over-Relaxation.

SPL Sound Pressure Level.

SQM Size-oriented Queue Management.

STE Short-Time Energy.

STMR Sidetone Masking Rate.

STZL Short-Time Zero-crossing Locations.

STZR Short-Time Zero-crossing Rate.

TCP Transfer Control Protocol.

TDMA Time Division Multiple Access.

TELR Talker Echo Loudness Rating.

TIA Telecommunications Industry Association.

UDP User Datagram Protocol.

UMTS Universal Mobile Telecommunication System.

UTAD University of Trás-os-Montes e Alto Douro.

VAD Voice Activity Detection.

VED Voice Enhancement Devices.

VoIP Voice over IP.

VoWiFi VoIP over WiFi.

VSELP Vector Sum Excited Linear Prediction.

WEPL Weighted Echo Path Loss.

WLAN Wireless LAN.

WSOLA Waveform Similarity Overlap-and-Add.

"In the beginning was the *word*..."

John 1, 1

1 Introduction

As far as we can go back to the history of Humanity, we realise that this is mostly a history of communication. Facts such as the *Homo Sapiens* emergent language, the Palaeolithic cave paintings (30 000 BC), the petroglyphs (10 000 BC) and the hieroglyphs (3 000 BC) and then the creation of the alphabets along with written languages, easily prove the primacy that communication has in the human existence. If a human being hopes to exist in a society or to get what he/she needs for progress and prosperity, he/she needs to communicate in an effective manner [1]. Back to the History, there are authors who advocate that the existence of a *Homo Sapiens*'s language was the basis of his survival when compared with his contemporaneous *Neanderthals* that died out [2]. Language is thus the heart of human life [3] and the *word* plays a preponderant role in it.

Among a multitude of means to communicate, the major medium of communication in a day-to-day experience is accepted as being the speech. Throughout History and across the world, people always used speech language to gossip, seduce, play, sing songs, tell histories, teach children, pray, pass on information, make deals, remember past, plan the future, *et cætera, et cætera* [3]. Therefore it is not a surprise to realise that innumerable authors and researchers payed so much attention to voice and speech.

In the late modern and contemporaneous eras, the need to communicate over long distances rose a new concept of communications –the Telecommunications. Despite the

history of telecommunications leads us centuries back to drums and smoke signals, telegraph, telephone and radio – that appeared in the 19th Century –, it was only in the last half century that Humanity has witnessed the telecommunications boom. Whilst radio and television played the main role on broadcasting news and entertainment audiovisual content for long decades, the telephone played the main role in voice conversations and so remained for longtime, seemingly with no end in sight.

Nevertheless, it was the widespread use of computers and the Internet that led to the telecommunications revolution as we are experiencing and benefiting today. In fact its ever increasing ubiquity and richness of technologies and protocols offer the flexibility that permits to transport and deliver a great diversity of media contents (multimedia) from anywhere to anywhere. As consequence, the telecommunications reality is nowadays concerned with a combination of exchanged text, speech, audio, still images (natural and synthetic), animation and video contents. Even though all these kind of contents form the current human communication language, none of them is so natural, and efficient, spontaneous, unambiguous as the speech language. Speech plays thus a primordial role in the telecommunications world [4]. However, since telecommunication systems are not ideal, the transmitted voice signals that represent the human speech are prone to errors and such errors affect the communication intelligibility. As the technology and services evolve, users tend to be less tolerant to failures and distortions. Therefore, along with speech communication technology evolution, there is also the constant need to enhance the services and quality experienced by users.

The research work developed in this thesis lies on this context by proposing methods to enhance the quality of voice signals and improve intelligibility of conversations when these are corrupted in transmission networks. This is the technological context where the work presented in this thesis was developed with the aim of enhancing the intelligibility, by means of signal quality monitoring and loss reconstruction.

1.1 Technological context

Voice communications have been evolving from the analogue Public Switched Telephone Network (PSTN) to the Integrated Services Digital Network (ISDN) and more recently to mobile networks mostly based on the circuit switching paradigm. On the other hand, the relatively recent packet-switching technology is responsible for the widespread use of the Internet Protocol (IP) that has reached traditional telephony communications in a global scale [5]. The Internet and its packet based architecture is becoming an increasingly ubiquitous communications resource, providing the necessary underlying support for many services and applications. As a consequence, Voice over IP (VoIP) is rapidly taking over legacy technologies supported by circuit switching, and the existing IP infrastructure soon appeared as a major transport mode for telephony voice signals because of its low cost and efficient utilisation of channel capacity associated with the flexibility, scalability and integration with of other Internet services as well as new value-added services that are possible to include [6–8]. Call centre integration, directory services over telephones, IP video conferencing, Fax over IP (FoIP) and Radio/TV broadcasting are among an ever increasing number of services and applications [9]. According to statistics from the Organization for Economic Co-operation and Development (OECD), the use of VoIP services has been steadily increasing in the last years [10]. For example, the French *Autorité de Régulation des Communications Électroniques et des Postes (ARCEP)* unveils that, from the third quarter of 2006 to the corresponding quarter of 2010, the use of the PSTN for VoIP services rose from 5% to 37% [11]. According to [12], it is expected that VoIP will replace conventional telephony in a couple of decades. Perspectives outlined by [13] state that the emerging VoIP traffic over the cellular network will represent 23% of the total voice traffic time by 2015 in occidental Europe. There is also a current evolution of voice services towards VoIP over WiFi (VoWiFi), which has recently emerged as a promising technology, and the bright prospects given by the emerging cognitive radio networks [14, 15]. With the new "Bring Your Own Device" (BYOD) trend it is also expected that the mobile phone is rapidly becoming our only phone and so a dramatic increase of using VoIP is also expected with the rise of the mobile VoIP applications. According to [16] a recent market research estimated a total number of 288 million mobile

VoIP users by the end of 2013. Today[1], in the United States, telephony communications of 40 percent of homes rely exclusively on mobile phones [17]. There are more users of cell phones in the world than Internet users and they also produce much higher revenues than Internet users [18]. A Juniper research has gone further on stating that there will be more than one billion of mobile VoIP users by 2017 [19].

Therefore, nowadays, VoIP is at the top of interest not only due to its technological importance but also due to the margin to growth that it still presents to current digital communication market. However, like all fast developing technologies and systems, there are still unsolved problems that strongly motivate researchers to continuously improve current technology performance and user experience.

Although there is margin for the VoIP market to grow, this is still conditioned by the quality experienced by users, which results in better service provisioning and eases market penetration [20, 21]. A known source of problems is due to the fact that IP technology was originally designed for data traffic and it only implements a best-effort network which does not ensure QoS because of the shared nature of available resources. As consequence, Internet data packets are prone to be lost. On the one hand, intensive traffic demands both high router load and high link load which causes that packets in the queues are dropped when traffic increases beyond a certain amount. On the other hand, collisions occurred due to excessive contention on the physical media access and the possible breakdowns in the transmission lines cause that packets become fragmented and so discarded. Moreover, the packet loss problem is still more considerable with the heterogeneous and cognitive radio networks, since they raise new challenges that run from new real-time constraints imposed by the new communication paradigm. For example the increase in the sensing periods is claimed to also increase the packet losses as well as the end-to-end delay and jitter (delay variation) which in turn increase the packet losses by discarding them [15].

For non-real-time traffic, the problem of packet losses is solved by the transport layer of the Transfer Control Protocol (TCP)/IP protocol suite that re-transmits lost packets. However, the delay introduced by the TCP protocol is normally too high for voice packets,

[1] 2014, Oct.

which will arrive too late to be useful. As solution, User Datagram Protocol (UDP) is used, instead. Although UDP does not have the heavy reliability mechanisms as TCP, it permits voice packet delivery with acceptable delay. However, with the lack of QoS mechanisms there is no way to impose a minimum delay bound neither to guarantee delivering of error free voice data to the receiver. For example, on the one hand, delays above 150 ms tend to become conversations unintelligible [22, 23] as well as delay variations above 100 ms [24]. On the other hand, if double-talks and mutual silence periods increase, then lower interactivity is obtained [25]. Thus, it is intuitive that such behaviour poses important problems to transport the data packets carrying conversational information.

In IP telephony, many impairment factors are identified as causing degradation on the quality that is experienced by the user, QoE, [24]: low bandwidth, excessive delays, jitter or lack of packet priorities. However, in a more or less scale, all of them contribute to packet losses. The constraints caused by low bandwidth may cause packet loss for a given period, thus not all packets might be able to reach the destination. Excessive delays imply that packets arrive out-of-the-time, making them to be useless when they arrive at the receiver and so discarded. Jitter makes that some packets arrive out-of-order and so discarded, too. It causes excess packet loss on receiving buffers, which depends on the the buffer size and the delay variance [26]. If different priorities are not used for transmission, then, in case of congestion, routers indistinctly discard both voice packets and data packets. Since voice packets are not retransmitted (*i.e.* UDP) these are lost, while data TCP packets can be recovered by retransmission.

When packets are lost, noticeable degradation of call quality occurs. Consequently VoIP must implement some mechanisms for handling lost packets. In order to maintain call quality, lost packets are substituted with interpolated data at the receiver. A practical use case was evaluated in [14], where the strong impact of packet losses in VoIP performance is minimised by a packet loss concealment strategy.

Overall, this is the technological context of the research work developed in the scope of this thesis, where the motivation is driven by the user QoE. To evaluate and improve the QoE in voice communication systems, is the main/overall motivation of this investigation

as addressed in the next section.

1.2 Research problems

It is known that, due to real time requirements, VoIP needs tighter delivery guarantees from the networking infrastructure than data transmission. While such requirements put strong bounds on maximum end-to-end delay, there is some tolerance to errors and packet losses in VoIP services providing that a minimum quality level is experienced by the users. Therefore, voice signals delivered over IP-based networks are likely to be affected by transmission errors and packet losses, leading to perceptually annoying communication impairments. Although sometimes it is not possible to fully recover the original voice signals from those received with errors and/or missing data, it is still possible to improve the quality delivered to users by using appropriate error concealment methods and controlling the QoS [27]. On the other hand, the evaluation of such quality is a key task so that the perceived quality can be assessed and thus appropriate actions on enhancing it can be performed.

Since the speech quality is a result of a perception and assessment process in which the subject establishes a relationship between the perceived and the expected auditory event, what makes users perceive a conversation as being of high or low quality is mainly the *a priori* expectation that they have about the conversation [28]. As the main quality reference has been for decades the PSTN service, users tend to compare the VoIP experienced quality with that of legacy PSTN network [14]. Concerning the voice service providers, it is thus of great importance that they have an accurate knowledge about the quality of their services [29]. It requires establishing QoS benchmarks so that the correct actions can be implemented to maintain and even enhance the service. Furthermore, these benchmarks have to be based on universally accepted metrics by using standard procedures to measure the quality. Voice service providers tend to measure the quality by means of classic objective measures such as delays and distortion. However, this kind of objective metrics does not represent the required accuracy as given by the true subjective users' opinions. More appropriate metrics, techniques and methods are necessary to effectively

measure the voice quality. In this context, chapter 3 addresses the voice quality evaluation problem and presents the most relevant voice quality standard evaluation methods.

Despite the fact that the human ear may not be sensible to small errors and packet losses of the voice signal, noticeable impairments are experienced when these factors attain a certain extension. In fact, glitches, artifacts, even silence periods may occur and perceived as annoying enough to make a conversation partially or totally unintelligible. To overcome this problem, error concealment techniques have been proposed [30]. Among these techniques, Time Scale Modification, Waveform Similarity Overlap-and-Add and even multicast techniques are used. They do not accurately recover the lost samples; they just conceal the effects. It is the human brain that "restores" the missing information by taking into account the perceived context. Some other recovering techniques have also been proposed. They include waveform substitution [31], generation of synthetic speech signal [32, 33], modeling of speech by taking into account the human vocal tract [34], embedding voice parameters in predecessor packets to reconstruct the lost ones, interleaving procedures [35] and pitch variation between the previous and the next known signal packets [36]. However, in general these solutions do not preserve the voice naturalness even creating voice artifacts. Therefore, it is important to derive techniques to recover missing samples while still preserving the speech naturalness. Chapter 4 deals with two linear interpolation algorithms that are useful for this purpose.

As previously mentioned, the best reliable voice quality measures are those obtained from real people evaluation when prompted to give their own opinion. However this apparently best practice is very onerous and time-consuming and cannot be performed in real time systems. A possible solution to overcome this problem is to use the standardised ITU-T computational methods, since they are reliable and some of them relatively easy to repeat. However a new constraint appears because the most widely accepted reference method (*i.e.*, PESQ) has protected Intellectual Property Rights if used for commercial purposes. Non-protected ones are not simple enough since they take into account impairment factors that are often out of the desired scope. A solution is to design or adapt a non-protected method such as the solution presented in chapter 5.

Due to the best-effort nature of the Internet where native QoS mechanisms are not available, voice packets are not guaranteed to be delivered in due time and so will be discarded in the receiver. Since data packets do not have the same urgency to be delivered as voice packets, QoS mechanisms may be implemented to give higher priority to voice packets. However, since this priority is given to data flows, all voice packets are considered as having the same importance without taking into account the individual importance of each packet for the global voice quality. A QoS scheme that takes into account the individual relevance of each voice packet is important to implement so that, in case of network congestion, less important packets are discarded whilst the most important ones are forwarded. This is the underlying idea that is explored in chapter 6.

1.3 Scientific contributions

The scientific contributions of this thesis are focused on the following four main fields.

1) **Signal reconstruction applied to VoIP.** The contribution to this field may be presented in four topics [37–40]:

   i) Study about the key parameters that influence the convergence of the linear interpolation algorithm used to reconstruct the observed signal by recovering the erased samples. Condition number, eigenvalues, matrix spectral radius as well as bandwidth of the signal were thoroughly studied and tested. With the achieved skills it is possible to control the reconstruction problem in order to put it in a well-conditioned point and thus make the signal reconstruction possible and efficient.

   ii) Implementation of maximum and minimum dimension linear interpolation algorithms, iterative and direct computation methods, simulation tests and comparison. Good results concerning accuracy and time efficiency have been achieved by using the minimum dimension algorithm which makes it suitable to signal reconstruction.

   iii) Based on the achieved skills, a combination of the proposed algorithm with a

proper packetisation scheme is proposed to reconstruct erased samples in a degraded VoIP signal.

iv) Contributions to an educational application (minimum dimension algorithm) to be used in signal processing as an aid tool to support teaching of signal interpolation related concepts allowing users to play with key parameters/issues and thus acquire skills in this area.

2) **Model for VoIP evaluation.** The contribution to this filed can be divided into two main topics:

i) The first topic concerns the research that was carried out by means of a thorough review about the most important methods available in the literature proposed to evaluate the voice quality in telecommunication systems. Special attention was paid to the standard methods proposed in the ITU-T recommendations under the design perspective [41, 42].

ii) An evaluation method for monitoring the voice quality in VoIP services was designed and implemented in a proprietary system of Portugal Telecom S.A. [43, 44][2].

3) **Voice packet classification.** Contributions to this field can be divided into two main topics:

i) Based on the evidence that different voice packets have difference importance on contributing to the global voice quality in a VoIP communication, an algorithm that optimally classifies packets according to their relevance was implemented. The underlying idea is that in case of network congestion the loss of the less important packets has lower impact on the voice quality than if losses occur randomly, as usually.

ii) Based on the packet classification algorithm, an experimental study was carried out to simulate the preferential discarding of the less important packets by giving them lowest priority. Perceptual voice quality was measured by the means of the MOS metric and compared with that obtained from random packet losses,

[2]Implementation was done by PTIn engineers.

simulating the real network behaviour. The results permit to propose the use of this classification algorithm to prioritise voice packets in order to enhance VoIP quality [45].

4) **MOS enhancement by combining packet prioritisation with linear interpolation signal reconstruction.** Based on the results and experience acquired with the carried out works using the signal reconstruction by linear interpolation algorithms and the voice packet prioritisation, a novel strategy that combines both these techniques was implemented. The results are seen as promising, since accuracy and computational effort enhancements were found when compared with the use of a linear interpolation reconstruction algorithm alone.

1.4 Thesis structure

This thesis comprises six chapters describing the research study carried out in this work and the different contributions, results and discussion.

In this introductory chapter the great importance of the human communication and specifically the speech, as well as the technological context and research problems were established as the main motivations to work in the voice processing field.

Chapter 2 presents the most important techniques that can be used to enhance VoIP. After describing the most important impairment factors that cause the degradation of the voice signal (and, hence, the voice quality), the chapter presents and discusses the most important and recent techniques available in the literature to enhance VoIP degraded signals. In this context, PLC techniques comprising both dissimulation of errors and recovering of lost parts of the signal are presented and discussed, as well as network techniques such as QoS enhancement and voice packets prioritisation according to their importance to perceived quality.

Chapter 3 presents a review of the most significant voice quality evaluation methods as they are described in the ITU-T recommendations. Classic metrics, such as the deciBel (dB) or the Signal-to-Noise Ratio (SNR) are mentioned to highlight that they do not

represent the true voice quality as evaluated by humans and the MOS is introduced necessary to validate voice quality evaluation methods, as stated in the ITU-T Rec. P.800. Standard methods are described and classified as subjective, objective and parametric; intrusive and non-intrusive.
Concerning subjective testing, the most procedural requirements are presented and subjective opinion scales for all methods are also explained. Two objective methods are presented: one of them using a reference signal while the other one does not need any reference signal. Special attention is paid to the computational PESQ method. Its nature and its relation with a subjective opinion scale as well as the technologies and applications for which this method demonstrates acceptable accuracy were given special attention, as stated in the ITU-T Rec. P.862 [46]. The "Single-ended method for objective speech quality assessment in narrow-band telephony applications", that do not need a reference signal, is also presented. Intrusive and non intrusive concepts are also defined.
Concerning parametric methods, the E-Model, as described in the ITU-T Rec. G.107, is presented [47]. The scope of this method is presented as well as the impairment factors that it takes into account. The reference model for the E-Model is depicted as well as the scope of its application.

Chapter 4 is devoted to voice signal reconstruction by using linear interpolation methods to recover lost samples. It comprises three parts.
The first part presents the algebraic fundamentals that are required to understand the linear interpolation methods related to the signal reconstruction by using systems of linear equations. Specifically, matrix notation, types of matrices, matrix characterisation, norms, eigenvectors and eigenvalues associated to eigenvectors as well as spectral radius and condition number are useful to characterise and condition the problems such that they can become well-conditioned.
The second part describes in detail two linear interpolation algorithms used to recover missing samples with the aim of applying them to voice signal reconstruction when VoIP packets are lost. The discrete version of the maximum dimension Papoulis-Gerchberg algorithm as well as a minimum dimension algorithm are described in detail.
In the third part of the chapter, the simulation results are presented and discussed. The influence of the spectral radius, error geometry and signal bandwidth on the problem

conditioning are studied. Also, a performance comparison between the maximum and the minimum dimension algorithms is addressed.

In chapter 5 a voice quality evaluation model is proposed in two modules. The first one applies to the analog circuit switching context and the second module applies to digital packet switching context. Both modules are modified versions of the E-Model, calibrated by using the PESQ as reference.
The module to evaluate the voice quality in the analog circuit switching context takes into account input factors such as attenuation, delay, echo and noise. The values concerning these factors were measured by using appropriated probes in the real circuit.
The module to evaluate the voice quality in the digital packet switching context takes into account the codec distortions and the packet loss rate that occurs in the real communication system by using appropriated probes in the circuit to measure it. It supports the G.711, G.729 and G.723.1 codecs and complies with the class C2 requirements of the ITU-T Rec. P. 564. It was integrated in Portugal Telecom Comunicações, S. A..

Based on the evidence that not all voice packets play an equal role on the voice quality, chapter 6 proposes an optimal packet classification algorithm aiming to assign high and low priority to VoIP packets to allow selective packet discarding when congestion occurs in priority networks.
This chapter comprises three parts. In the first part the classification algorithm is described. In the second part, simulated results using this algorithm are presented and discussed. Packet losses are simulated in erasure channels by taking into account the priorities assigned by the classification algorithm and using two random processes: Bernoulli and Gilbert models. In the third part of this chapter, the classification algorithm is combined with the Papoulis-Gerchberg (PG) algorithm, as described in chapter 4, as a means to lever the performance of reconstruction when the PG algorithm is used.

Finally, chapter 7 presents the conclusions and discusses several topics for future work.

2 Techniques for enhancing voice quality

This chapter presents a review of the most important techniques currently available in the literature to enhance VoIP in networks where a great diversity of impairment factors cause that voice quality and conversation intelligibility become degraded. In the introduction the major factors contributing for the impairments are presented. In section 2.2 the most recent and important techniques to enhance VoIP are presented and discussed.

2.1 Introduction

Despite the significant evolution of voice communication systems in the last years, modern systems still cannot avoid degradation of the intelligibility experienced by users of such systems. In fact, disturbances such as distortion, delay, packet delay variation (*aka* jitter), packet loss, attenuation and echo are among the most important impairment factors that psychologically affect the quality provided by a voice communication system.

The following three categories of impairment factors are identified as the major contributors for reducing the quality, due to the difficulties they cause to users: i) factors that lead to **listening difficulty**; ii) factors that lead to **talking difficulty** and iii) factors that lead to **conversational difficulty** [28, 48].

Listening difficulty

By "listening difficulty" it is meant the difficulty that listeners have to understand what they are hearing, due to the distortion in the voice signal. Among the diverse factors that cause such distortion, two important ones have to be taken into account:

- Transmission errors – This kind of impairment is related to unreliable transmission channels, where noise and interferences are major factors.
- Data losses – This is primarily related to the packet losses due to network constraints. Since Internet is primarily a best-effort network, delivery of real-time traffic in due time, such as voice traffic, is not ensured. For example, in the case of network congestion, several disturbing factors may occur, like excessive delays that make the packet contents useless when they arrive at the receiver, or excessive delay variations that cause the received message to be cropped in time. Routers indistinctly discard packets as it is momentary needed to cope with bandwidth limitations. As result, packet erasures occur where voice signal is expected at the receivers and thus degradation of the voice quality is experienced by the end user.

Talking difficulty

By "talking difficulty" it is meant the effort that talkers have to make to be understood. Talker echo and incorrect sidetone or room noise are the main contributing factors:

- Talker echo – This kind of impairment concerns the portion of the talker voice that is returned with sufficient delay to distinguish it from the original voice: it causes discomfort on talking.
- Sidetone – This occurs when the transmitted sound goes from the microphone to the phone receiver of the same telephone set. In this case the delay is negligible but there may be loudness issues that annoy the talker. On the one hand, the absence of sidetone causes discomfort since it makes talkers think they are not being heard; on the other hand, too much intense sidetones may lead talkers to keep away from the handset which in turn may cause their voice not to be well captured.
- Room noise – This kind of impairment consists on the external sound perturbation that talker experiments, forcing him to annoying effort in order to be understood.

Conversation difficulty

"Conversation difficulty" depends on the discipline effort that subjects have to do in order to not talk at the same time as his counterpart. This is specially critical when excessive transmission delays make listeners think that talker is quiet. The main factor that contributes to degrade the conversational quality is delay. It is well known that long delays, as for example those experienced in satellite links, lead to continuous hesitations by the speakers, which are very annoying and thus make reduced users' QoE.

To cope with this harmful reality and attain an acceptable voice quality, techniques to enhance and make the transmitted signal more robust must be taken into account. In section 2.2 the most significant techniques for enhancing VoIP are presented. They comprise two main aspects needed for achieving such aim: by recovering or dissimulating the lost parts of the voice signal and by using network techniques such as QoS enhancement and packet priorities. Subsection 2.2.1 and subsection 2.2.2 deal with these two aspects, respectively.

2.2 Techniques for enhancing VoIP

As previously referred to, transmission errors in voice communications, and particularly in voice over IP networks, are known to have several different causes but the single effect of delivering poor quality of service to users of such services and applications. In general this is due to missing/lost samples in the signal delivered to the receiver.

Channel coding can be used to protect transmitted signals from packet loss but it introduces extra redundancy and still does not guarantee error-free delivery [49]. In order to achieve higher quality in VoIP services with low delay, effective enhancing techniques must be used at the receiver.

Packet losses may not be totally harmful, since small losses are imperceptible to the human ear. However, beyond a certain loss rate, they tend to become harmful to the voice intelligibility [50]. Among the multiple techniques used to alleviate this problem, QoS enhancement, PLC, packet recovering and packet prioritisation are the most relevant, addressed in the next subsections.

2.2.1 Packet loss concealment

To mitigate the effects of packet loss, different techniques can be used to recover or dissimulate the lost parts of the voice signals, based on the nature of human hearing, which is prone to ignore minor mismatches in short periods and restore them according to the context [51]. Techniques to deal with data loss may be classified as reconstruction techniques or PLC techniques. Whilst the former try to restore the original samples that are missing, PLC techniques mainly try to hide transmission losses to a certain extent. In this case it is the brain that reconstructs the remaining missing contents.

The objective of PLC may also be to generate a synthetic speech signal to substitute missing data (erasures) in received bit streams. Since speech signals are often locally stationary, it is possible to use the signals' past history to generate a reasonable approximation of the missing segment. If the erasures are not too long, and not located in a region where the signal is rapidly changing, the erasures may be inaudible after concealment [33].

Among several possible solutions for recovering or concealing lost packets, it is worth to mention those algorithms that try to reconstruct the missing segment of the signal from correctly received samples [33, 38, 52, 53].

Zero amplitude stuffing is a rudimentary technique that consists of setting all missing packet samples to zero. It is the easiest way to deal with the problem but has no useful effect and may, even, be worst than if comfort noise was added instead of silence [54]. It is mentioned as one of the first techniques just for historic purposes.

Waveform substitution is a method to replace the missing part of the signal with samples of the same value as its past or future neighbours [31]. Waveform substitution is, perhaps, one of the first and simplest used techniques, but, also rudimentary. It requires very low computational resources while giving relatively good results. Waveform substitution often results in unnatural, "robotic" sound when a long burst of packets is lost and, when not associated with other technique, the results may be worse than if this technique was not used because it introduces discontinuities at the packet edges. Furthermore, the replaced

segment may be not enough correlated with the replacing one, since no special similarity is taken into account when simply using the precedent segment. In [55] a double sided periodic substitution method for recovering missing samples is presented. However, since it is based on signal substitution, it suffers from clicking noise [56].

In [56] a method is proposed to restore clipped audio signals, by using Recursive Vector Projection technique. The estimated signal is said to hold consistency with neighbouring samples so that the restored signal does not suffer from click noise. This technique is claimed to improve SNR but when related to subjective assessment no noticeable difference was found.

The problem of the voice signal clipping is also addressed in [57] where a method is proposed to enhance clipped voice signal by restoring it. An improvement method as well as nonuniform normalisation and quantisation method are used and applied by detecting signal peaks and valleys. It is not known how effective the method is since authors do not present any measure of achieved enhancements.

The technique of pattern matching, presented in [31], intends to overcome the lack of similarity previously referred to, between replacing and replaced segments. When the k^{th} packet of L samples is lost, the last M samples of the $(k-1)^{\text{th}}$ packet ($M < L$) are used as a template in a search for a substitution packet. The template slides along a finite search window until the best match is found. The substitution segment is created with the L samples immediately following the best match to the template. The idea is that if these two sequences of length M match, then so should the subsequent L samples (one packet length). Despite the probability that these L samples are correlated with the missing packet due to the stationary nature of the voice, when the voice presents rapid transitions, this method is likely to fail. Moreover, this technique requires considerable computational delay. The maximum tolerable packet loss rate is 20%.

Methods with increased error robustness embed voice parameters of a packet k in its predecessor $(k-1)^{\text{th}}$ or in the subsequent packet are proposed in [58, 59]. Based on this idea, two reconstruction techniques are proposed in [58] that send in the packet $(k-1)$

two slow-varying parameters of speech packet k: the Short-Time Energy (STE) and, depending on the method, either the Short-Time Zero-crossing Rate (STZR) or Short-Time Zero-crossing Locations (STZL). Using these methods, amplitude and frequency continuity between the concealed waveform and the lost one are said to be ensured. In case of losing packet STE and STZR/STZL, the parameters are used to reconstruct the waveform to replace the lost packet. The computation of these parameters involve considerable delay. These methods may be used when the natural speech silences exist to permit to proceed with such computation; otherwise it is likely that not enough time is available to perform such calculations.

Some other reconstruction techniques use voice related parameters, such as the pitch of the adjacent frames of the missing frame (also called "gap" in some literature) in order to reconstruct the respective waveform [36, 59]. For a mix of single, double and triple frame losses, MOS is claimed to vary from 4 to about 3 when frame loss rate varies from 0.5% to 10%, respectively [36].

In [60] a method to address the problem of interpolating the missing data over short gaps in audio signals is presented. The problem is formulated in terms of Gabor regression model in which information from the surrounding time-frequency plane is leveraged and an interpolated signal is obtained by keeping the qualities of the signal under consideration. In two examples where about 35% of samples are missing, gains of about 10 dB and 6 dB are achieved. The method requires considerable computational resources.

Interleaving procedures at the packetisation stage are, also, considered. Splitting the even and odd samples into different packets is another method which eases interpolation of the missing samples in case of packet loss. Since original contiguous samples are broken apart, a packet loss does not represent an entire temporal segment to be lost, but only a few of its samples which facilitates the use of reconstruction techniques at the receive side [35, 61].

Voice synthesis is an alternative technique by which missing voice is synthesised from the previous known signal [32]. However, this kind of replacement tends to not render voice as natural as human voice.

In [34] a method to replace missing packets that models the speech by taking into account the human vocal tract is proposed. It uses a two-side approach since it models the signal that precedes the gap and also the signal that succeeds the gap. This approach is claimed to have a lower Mean Square Error (MSE) when compared with the ITU-T G.711 Appendix I PLC method [33]. The method requires that the packet lengths should be sufficiently long in order to accurately calculate the needed parameters.

The different approaches to deal with voice communication errors can be classified in either source-coder independent or source-coder dependent [62]. The former schemes implement loss recovering methods only at the receiver, where concealment techniques are included. In such receiver-based schemes, the effects of the packet losses may be dissimulated by using signal reconstruction algorithms. Particularly interesting to the work presented in this thesis are the discrete version of the Papoulis-Gerchberg interpolation and the minimum dimension interpolation algorithms, since they permit to reconstruct voice signal with good accuracy, while preserving the original voice naturalness and requiring relatively low computational resources [38, 63].

The source-coder dependent schemes might be more effective but also more complex and in general higher transmission bandwidth is necessary: the sender first processes the input signals, extract the features of speech, and transmit them to the receiver along with the voice signal itself. For instance, in [64] the authors propose to use additional redundant information to ease concealment of lost packets.

This question is so important that many of the standard Code-Excited Linear Prediction (CELP)-based speech decoders have PLC algorithms built into their standards. When the receiver detects a loss, it triggers concealment procedures that take into account the error-free packets received before the lost one(s), even the packets after the lost one(s), and try to synthesise voice to cover the gap created by the lost one(s).

In [65] a frame erasure concealment method is proposed to apply to a CELP codec –the Adaptive Multi-Rate Wideband (AMR-WB) codec. It combines the constrained Adaptive Codebook (ACB) contribution at the encoder with the ACB resynchronisation at the decoder and propose an improved frame erasure concealment algorithm based on parameter

re-estimation of the pitch and gains of ACB by using linear prediction on the decoder and the Innovative Codebook (ICB), making use of both previous and future frames. The best results occur when five future (and past) subframes of 5 ms are used, which implies a delay of, at least, $(5+1) \times 5$ ms $= 30$ ms. MOS enhancements are presented for Frame Error Rates ranging from 1% to 10% and vary from about 1.9 to 1.1, respectively.

In G.729, an error concealment procedure is incorporated in the decoder to reduce the degradation in the reconstructed speech due to frame erasures in the bit stream. The coder is more robust against random bit errors by computing and transmitting a parity bit. In case of mismatch in the parity bit or when a frame is identified as erased, the concealment procedure is applied. The pitch analysis algorithm exploits only current frame information to avoid multiple pitch lags [66]. This algorithm can be improved by taking into account the previous frame as it occurs in the G.729.1 codec [67].

In G.729.1 ("An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729"), redundancy is added to the bitstream by determining and transmitting few supplementary concealment/recovery parameters such as signal classification information, energy information and phase information (the estimated position of the last glottal pulse in the previous superframe), in order to proceed with frame erasure concealment by the decoder [67].

In G.723.1, an error concealment strategy for frame erasures is included in the decoder. If a frame erasure has occurred, the decoder switches from regular decoding to frame erasure concealment mode [68].

The algorithm described in Appendix I of the ITU-T Rec. G.711 describes a concealment method for the G.711 system. A copy of decoded output is continuously being analysed so that the pitch period can be calculated and waveforms are extracted. When an erasure is detected, this information is used to synthesise the needed signal [33].

Most of the source-coder independent techniques are based on signal reconstruction algorithms which use interpolation techniques combined with packetisation schemes at the source that help to recover the missing samples of the signal [61, 69].

Packet Loss Concealment algorithms for CELP-type speech coders are usually based on speech correlation to reconstruct the speech of lost frames by using parameter information obtained from the previous correctly received frames. However, these PLC algorithms have difficulty in reconstructing voice onset signals since the needed parameters are mostly related to silent frames because when the current frame is a voice onset, several previous frames could be silent (or noise) frames [54].

In [54] it is proposed a receiver-based PLC algorithm for a CELP-type speech coder to improve the performance of speech quality when frame erasures or packet losses occur in wireless sensor networks under burst packet loss conditions. It is based on a multiple codebook-based approach that includes a traditional adaptive codebook and a new random codebook composed of comfort noise. The achieved MOS enhancements are relative to the native G.729-PLC algorithm. They range from 0 to about 0.8 for single packet loss rates varying from 0 to 8%, respectively, and ranging from 0.05 to 0.12 when the loss bursts vary from 1 to 3, respectively. The computation requirements are relatively significant.

Among the most used methods are those based on Time Scale Modification (TSM) of audio signals, by which the duration of the signal is modified while maintaining other characteristics such as the pitch. Typically such techniques extract features from the received signal and use them to recover the lost data.

Another kind of methods that is commonly used in real time voice communications is the Waveform Similarity Overlap-and-Add (WSOLA) and its enhancements [51, 70–74]. By using this technique, when a lost packet or frame of the original signal is detected, the duration of some known packets or frames before the lost one(s) is slightly extended in order to cover the gap left by the loss one(s).

In [72] perceptual differences between different signal sections are recognised and used on proposing enhancements in the classic WSOLA method used to conceal missing sections in a signal. To do this, perceptual significant transients are detected and so leaved intact, whereas remaining steady state signal is subject to the WSOLA inherent time-scale modification.

What these WSOLA techniques make is to ensure a certain continuity in the signal but does not recover the lost packet contents; it just conceals the problem. It is the human brain that restores the waveform according to the current context [51].

An adaptive delay concealment method that exploits the time-scale modification scheme is presented in [75]. By applying a variable degree of stretching for each packet, each one could contribute to adapting to the network delay as well as packet loss. The proposed algorithm results in 1.5% to 8% improvement over the reference [76]. This relies on a content analysis performed in the sender to classify the input speech as silence, transient or general segments by using short-time energy and zero crossing rate. At the receiver, a set of timing parameters are used as well as the playout length. The resulting audio quality is said to exhibit better intelligibility, despite the artifacts introduced by the stretching procedures. However, the algorithm complexity is considerable.

In [30] different strategies for improving voice quality over Wireless LAN (WLAN) are considered. They rely on the use of short term linear prediction to achieve "sample loss concealment" instead of the standard PLC strategies that consider concealment of the entire packet. Contrary to the IEEE 802.11 standard, which discards packets whenever even a simple bit is corrupted, the proposal is to process Media Access Control Protocol Data Units (*aka* frames), whose error bits are located in the most significant bits of the payload whilst discarding those Protocol Data Units whose error bits are located in the control bits (addresses, among others). A reduction of 80% in the packet loss rate is claimed for low bit-error rates. The proposed method requires changes in the standard, namely the adding of 16-CRC bits and 160 parity bits in the payload.

A "Multicasting Routing Algorithm" is proposed in [77] to enhance QoS of VoIP over heterogeneous networks (WLAN and General Packet Radio Service (GPRS)). Such enhancement is achieved by improving handoff speed and throughput. In particular, packet loss is claimed to be completely eradicated by multicasting the packets during the handoff period and switch back to normal mode after the handoff. Handoff latency is said to be reduced in 18% from its normal value. From the graphics it is possible to infer a packet loss reduction of 80% approximately. With this method, the throughput increases during

the handoff period and considerable bandwidth margin is needed, due to multicast.

Playout scheduling schemes were also proposed to conceal late losses [71, 78–81]. Nevertheless additional delays are introduced that are proportional to the depth of late losses to recover, and discarded packet effects are not minimised. In [82], Kim *et al.* use a combination of an adaptive playout scheduling and a packet loss concealment technique by using effective signal classification to enhance VoIP speech quality at the receive side of a mobile Internet phone. The results show considerable enhancements in regard to packet losses, buffer delays, jitter estimation errors and PESQ scores when compared with three reference methods. The proposed method uses several algorithms needed for packet classification, adaptive packet loss concealment, signal merging, adaptive playout scheduling and network jitter estimation. Despite the lower computational demands relative to the three references, it still is of great computational complexity.

In next subsection other kind of techniques for enhancing VoIP is presented that rely on improving the classic QoS parameters while distinguishing the role that each voice packet has on the final voice quality.

2.2.2 QoS Enhancement and packet prioritisation

One of the strategies to improve VoIP quality is by means of enhancing QoS through important parameters and/or giving higher priority to some voice packets than to others or to those voice packets that are the most representative for the conversation intelligibility.

A "Proactive QoS Enhancement Technique for efficient VoIP performance over WLAN and Cognitive Radio Network (CRN)" is proposed in [83]. It uses an optimisation algorithm that proactively configures codec parameters to keep loss and latency within tolerable limits. For example, by decreasing the *packet per second* parameter of codecs, latency and packet loss reduction are achieved since Access Points buffer overflow is mitigated. An active queue management that uses Random Early Detection (RED) discipline is implemented so that codec parameters are proactively configured. By proactive it is meant that RED ensures that codec parameters adjustment (as for example the bitrate) is performed even before the queue is full. The results show QoS enhancements concerning

throughput, packet loss and delays. The drawback is that the algorithms that constitute the proposed technique imply changes in the sensing and transmitting intervals relative to the standard CRN.

In [84] enhancement of VoIP is achieved by improving QoS in Wireless VoIP networks. The strategy takes into account the two main situations in which Wireless VoIP is used: i) when the user is stationary and it is the background noise that degrades the speech quality and ii) when the user is moving with a VoIP phone and channel interference, fading or real world interference (engines, RF interferences) are the degradation factors. To achieve VoIP betterments in such situations, the authors proposed a recursive filter using Kalman filtering that combines both stationary and dynamic effects for speech enhancement in a wireless VoIP environment. Presented results concern noise reduction and, as it is shown, MSE decreases from 0.06 to 0.013 after 200 iterations of the Kalman filter run.

In [85], an adaptive packet scheduling algorithm which serves different traffic queues based on QoS levels of each traffic queue and variation of available spectrums is proposed for QoS maintenance and enhancing in real-time traffic, to be applied in heterogeneous Cognitive Radio Networks. Priorities are given as a per queue and per user basis, where three kind of weights for balancing impacts of delay and throughput distinguish VoIP, Moving Picture Experts Group (MPEG) and FTP traffics according to the intended priorities. Moreover, a channel-adaptive coefficient influences priorities by inclining to real-time traffics when the available sub-channel number decreases. The authors refer that the proposed algorithm provides better QoS guarantee for real-time traffics (as a whole). However, by using this algorithm for the specific case of VoIP traffic, the presented average packet delays are always worse than the three other reference algorithms. In regard to the cumulative distribution of packet drop rate, the results show that the probability to lose VoIP packets is greater than other traffic types or methods, since cumulative distribution of VoIP packet loss rate have greater margin to grow above 0% of packet drop rate than other types of algorithm and/or traffic types.

In [15] modifications are proposed in the "cognitive radio cycle" to improve QoS in Cognitive Radio Networks. Sensing to locate unused spectrum segments is proposed to be

performed, as usual in the sensing period, but now also between successive transmissions, contrary to the usual separate sensing and transmission intervals. This practice reduces end-to-end delay, jitter and packet loss QoS parameter values. Furthermore by sending more than a single packet in one transmission slot before sensing the channel, the authors claim that packet loss is further reduced. The proposed approach requires that changes are made in the established CRN protocols.

The work of [23] aims to enhance VoIP quality in WLANs by doing an appropriate selection of the routing protocol in an ad-hoc IEEE 802.11 environment. By analysing the performance of the Dynamic Source Routing (DSR) protocol [86], GPS-Based Addressing and Routing protocol (GRP) [87], Ad-hoc On-Demand Distance Vector (AODV) [88] and Optimised Link State Routing (OLSR) [89] routing protocols, an order of merit is established in which OLSR is the protocol that better optimises the QoS parameters under study: it maximises the throughput and minimises delay. DSR and GRP are *ex-æquo* in the last position. The results assume a data rate of 54 Mbps, G.711 codec and adaptation of the OLSR protocol.

In [90] generic VoIP traffic is differentiated from other sources and thus protected by giving it major priority. Furthermore an individual-based prioritisation scheme (Drop Sel with Delay) is proposed in which aged voice packets are discarded in intermediate routers in case of network congestion. Since these packets would be late and thus not usable at the final receiver, opportune drop implicitly results in giving more priority to useful packets while conserving bandwidth. In this sense, discarded packets are those of less perceptual importance since when they would arrive, its perceptual importance would be null, simply because they would arrive out of the time. This strategy do improve QoS by increasing throughput and mitigating loss rate. Despite the presented MOS enhancement as a whole in the voice samples, once a voice packet is lost, the consequent voice erasure does not permit recovering neither concealment, which results in clearly perceptible impairments.

In [91] QoS enhancement in a femtocell network is claimed by proposing a priority queuing policy that avoids data losses resulting from congestion traffic. It combines priority classes with adaptive time allocated for transmissions as a function of the spectrum availability.

As consequence, packets will only be transmitted when there is a slot for transmission and they are in the right queue of priority as given by their traffic type. Low-priority packets are buffered whilst the highest priority packets are sent. Throughput is claimed to be increased and end-to-end as well as medium access delays, decreased. However, the specificity of each packet on the final voice quality is not taken into consideration and the quality experienced by the users is not considered.

A Size-oriented Queue Management (SQM) scheme that implements two collaborative mechanisms is used to favour time-sensitive traffic such as VoIP [92]. A dropping mechanism is used to classify incoming packets into traffic classes and hence assign different dropping probabilities. It relies on the moving average of packet sizes. Small packets have smaller dropping probabilities. A scheduling mechanism rearranges packets in order to grant faster service times by promoting time-sensitive traffic. This size-oriented strategy to determine which packets must be prioritised is based on a statistical dynamic correlation of packet sizes: a moving average of packet size is used to determine if the arriving packet at the router is a big or a small packet. Despite the augmented user-perceived quality, probabilities given by classes of traffic lack some service granularity (perceptual precision) since this is not absolutely representative of the perceptual relevance of each packet to the overall voice quality.

Assuming that VoIP packets are of small size, the authors in [93] propose a service differentiation for small packets in which they dynamically get some limited priority over long packets. This limited priority comes upon the impact they have on long packets since if this limitation was not implemented, cumulative delays would deteriorate too much the QoS relative to non-prioritised long packets. This type of priority mechanism is equivalent to assign high probability to VoIP packets by promoting small packets in the queue. In [92] and [94] the same criterion of packet size is used to distinguish time-sensitive flows. Despite the achieved service granularity, this differentiation is based on the correlation that exists between small and VoIP packets. This granularity does not take into account the perceptual importance of each packet on the global voice quality perception.

Some other packet prioritisation approaches have been proposed to improve the QoE

in voice communications. In [6] a voice packet priority protocol based on the Older-Customer-First (OCF) and Earliest-Deadline-First (EDF) priority disciplines is proposed to dynamically change the priority of each packet. This change is mainly based on its age, where the given priority is proportional to its age in order to minimise the variability of packet delay. The jitter is thus minimised which permits to better allow for a reliable signal reconstruction. The proposed model is mainly focused on the jitter minimisation and, as stated in the conclusions, it is most effective for long routes and heavy traffic, when delay variability is most likely to be significant. The increased processing and sorting necessary in each queue could be too costly or create additional delay, impairing the achieved effectiveness.

Some of the previous prioritisation approaches are mainly based on priority assignment to classes of service or classes of priority that are determined by the type of flow, in a per-queue or in a per-user basis where the traffic type or packet size constitute criteria. However, in each flow, the specific perceptual contribute of each packet is not taken into consideration. Although some of the approaches are packet-based, they do not consider the intrinsic features of voice quality.

Based on the observation that not all voice packets equally contribute to the final voice quality, a classification scheme that takes into account individual voice packet contribution was recently proposed in order to give more priority to the most perceptually important packets [95]. For example, packets containing transitions between unvoiced and voiced speech are perceptually more important than a single packet within a sequence of packets representing a stationary voiced segment [96]. By giving more priority to those packets, quality in VoIP can be improved when packets have to be discarded. By further using a PLC method, it is more likely to conceal a packet loss within a stationary segment rather than in a transition segment. This kind of classification can be of primordial importance to further increase the concealment effectiveness.

In [7] it is shown that the loss of voiced frames after an unvoiced/voiced transition leads to a significant degradation of the speech quality while the loss of other frames is concealed rather well by the decoder's concealment algorithm. The authors have developed

a selective packet prioritisation scheme that protects those packets which are essential to the speech quality by marking them at the sender with a higher Differentiated Services (DiffServ) priority, while relying on the decoder's concealment in case other low priority packets are lost. High priority is assigned to those frames that contain either voiced and unvoiced-to-voiced transition voice segments. Unvoiced segments are assigned low priority. The priorities are applied at the wireless link layer to identify and retransmit only these essential packets. The results led the authors to suggest a mix solution to improve voice quality that includes packet prioritisation and retransmission of high priority packets which retransmission rate must increase with the Bit Error Rate (BER) (typically close to 10^{-3}). Beyond the delay caused by the prioritising task, the retransmission strategy of this solution increases the congestion probability when BER is high and causes further delays. Moreover, the voiced/unvoiced classification is not so related to voice quality as if the classification would rely on the packet contribution to the final MOS.

In [96] the unequal perceptual importance of voice packets is the motivation to propose unequal error protection methods that allocate more error-control resources to certain packets over others. Perceptually critical voice packets are provided with greater error protection. In particular, protection scenarios include varying the number of copies of a packet that are piggybacked onto subsequent packets and adaptive Reed-Solomon (RS) Forward Error Correction (FEC) scheme in which only certain packets are provided with RS-FEC protection. For this, sender anticipates what the PLC will do in case of packet loss, calculates the expected distortion for various protection scenarios and selects the optimal protection policy by using Lagrangian optimisation. The authors present a MOS enhancement of 0.2 to 0.3 when comparing their methods with those of equal error protection schemes. The proposed methods obviously increase the necessary bandwidth by piggybacking until three packets and introduces considerable delay on performing all the needed computation at the sender plus waiting for to the 3rd packet to recover the lost one.

In [97] a "Speech property based booster" is proposed to improve quality of voice over Wireless Local Area Network (LAN). Differentiation of voice packets is performed according to their importance for perceptual quality. Results coming from this application-level

differentiation are directly mapped to the data-link level, which permits to classify this approach as a transparent "protocol booster" approach since IP packets do not need to be modified. Important packets are protected at the data-link layer by three mechanisms: retransmission of high importance packets, redundant transmission and a combination of both. The results show enhancements in the frame loss rate as well as in the perceptual distortion. The drawback is that both retransmission and redundant packets demand more bandwidth. The referred transparency sounds good but on relying the boost process on the data-link layer it limits its application of such a scheme to the first hop [7].

Despite the fact that the above techniques increase classification granularity, the most important criterion to classify packets would rely on the contribution of each packet to the subjective quality (MOS) of the final voice. A method based on differentiating the contribute of each packet to the voice quality is herein proposed in chapter 6.

2.2.3 Other techniques

Other type of techniques are based on the traditional FEC [98, 99]. Redundant information is added to voice packets to reduce the effects of packet loss. However, these would introduce significant extra delay and increase the amount of used bandwidth [54]. For this reason there are who claims that they are not appropriated for real-time communications [100].

As alternative to the FEC technique, Multiple Description Coding (MDC) technique may be considered. By using it, a certain amount of redundancy is introduced by splitting the bitstream into multiple streams or paths, which means that in case of loss of one of them, the voice signal can still be intelligible despite its lower quality. The drawback is that this technique consumes a wider bandwidth [54].

In [8] and [101] a dual stream approach to mobile VoIP, in which the single VoIP flow is duplicated and sent through two different mobile operators is presented. An average packet loss reduction of 60% as well as a MOS improvement of ≈ 0.8 are claimed. In terms of cost/benefits, it is possible to see that it is needed to duplicate the bandwidth

–since each packet is sent twice– to reduce packet loss in 60%. Furthermore, this approach requires the use of devices provided with dual Radio Frequency (RF) modules and mechanisms that implement load balancing.

2.3 Conclusions

In this chapter the most recent and significant techniques aiming to enhance the VoIP quality were studied and presented by focusing on the three principal current strategies: Packet Loss Concealment and recovering, QoS enhancement and packet prioritisation.

Some primitive techniques like "zero amplitude stuffing" and "waveform substitution" were mentioned so that the problem can be framed and to provide the basic understand for the subsequently presented techniques. Other techniques include pattern matching in which voice characteristics of packets adjacent to the missed one are taken into account when recovering the lost packet. The embedding of packet voice parameters in their predecessors was also referred to as well as interleaving procedures and voice synthesis. Concealment techniques inherent to waveform and CELP codecs were also addressed. Other techniques as Time Scale Modification and Waveform Similarity Overlap-and-Add are also reviewed.

Strategies contributing to QoS enhancement include decreasing packet loss, latency and jitter by means of changing codec parameters, achieving packet redundancy by duplicating and retransmitting them. Differentiation of traffic and establishment of priorities were also presented as well as some per-packet priority variants.

All these techniques aim to enhance the perceived voice quality which must unequivocally and accurately be measured so that the quality of voice communication systems can effectively be referenced, as well as benchmarks and service level agreements established. In line with this kind of needs, next chapter presents the most important voice quality evaluation methods currently in use.

3 Methods and metrics for evaluating voice quality

This chapter presents a review of the most important voice quality evaluation methods that are recognised as necessary background knowledge to the work presented in this thesis. In the following sections, the most relevant aspects and the different methods related to telephony voice quality evaluation are described. Subjective and objective methods are addressed as defined in standard ITU-T recommendations and the E-Model is also described as the most important parameter model in the scope of this research.

3.1 Introduction

Measuring the voice quality at different points along the communication chain is necessary not only to find where the main quality constraints are imposed but also to improve the performance of services and applications provided to users. The success of new voice technology or any other service largely depends on the user-opinion of the perceived quality. Thus it is of crucial importance for developers and service providers to assess the voice quality either using off-line mechanisms or real-time evaluation for continuous monitoring and performance analysis.

Despite the potential problems, VoIP services, for example, should be able to achieve comparable voice quality to that of legacy PSTN networks in order to be competitive [14].

Although VoIP services often offer much cheaper solutions than PSTN, what really matters is essentially the user perception of the service quality. Therefore, it is of great importance for service providers to establish QoS benchmarks of their services in order to have universal metrics and procedures that quantify how users experience such quality, *i.e.*, the **Quality of Experience** (QoE). The importance of QoE in voice communication has been increasing in the last years for both the research community and industry.

A great number of quality evaluation methods have been proposed in the literature and applied in diverse application contexts. For example some of them are used to evaluate the quality of synthesised voice ([102, 103]), others in the healthcare area ([104–106]), voice coders ([107–109]), packet voice systems ([110–114]) or even to assess acoustic information [115]. Among the vast research work done in this area, some of them assume special importance, since they resulted in relevant contributions to ITU-T standard methods and procedures. This is the case, for instance, of the works described in [116] and [117], that contributed to the ITU-T Rec. P.800 [118], the works described in [119], [120] and [121], that contributed to the ITU-T Rec. P.862 [46] or the works described in [122] and [123], that contributed to the ITU-T Rec. P.863 [124].

In the context of voice quality evaluation, the ITU, through its Telecommunication Standardisation Sector (ITU-T), has released a set of recommendations to standardise metrics and methods to carry out proper evaluation of telephony voice quality. These methods take into account the most significant human voice and audition characteristics along with possible impairments introduced by current voice communication systems that are reflected in the voice signal, such as noise, delay or distortion due to low bitrate codecs, transmission errors and packet losses.

3.2 Objective metrics and concepts

The most classical and perhaps the easiest way to evaluate voice quality relies on the computation of objective metrics that take into account the errors and distortions associated to a particular signal. Such metrics include parameters such as the SNR or the Segmental Signal-to-Noise Ratio (SNRseg), the decibel (dB), the Peak Signal to Noise

Ratio (PSNR), the signal level, the MSE, the Root Mean Square Error (RMSE), delay, delay variation, packet loss rate, inverse linear unweighted distance, unweighted delta form, energy ratio or cepstral distance. Each of them present different accuracies in the evaluation of voice quality as perceived by users [114, 125].

Although these metrics seem to be attractive due to their easy computation and objectiveness, they do not provide, *de per si*, a reliable evaluation metric about perceptual factors, such as intelligibility. Classical quality measurement techniques using concepts such as SNR and frequency response functions have become grossly inaccurate [126]. In fact the actual perception of voice quality varies from person to person, including variable psychological and cognitive factors, such as personal expectations, content of the conversation, tolerance to accept noticeable artifacts and the subject humour. It is widely recognised that conventional metrics such as frequency response or SNR ratio cannot reliably assess the quality of such systems [111].

A key factor affecting the subjective quality is the expectation of the listener because when a speech signal reaches the human auditory system it establishes a relationship between the perceived and the expected auditory event. The speech quality that is actually perceived is thus a result of the human perceptual system combined with some subjective assessment process [28]. It also depends on the evaluator subject, hear reliability and even auditory tastes. For the same subject, the perceived quality even varies along the time. Therefore, to be sufficiently reliable, voice quality assessment must take into account its inherent subjective nature. Subjective evaluation gives the most reliable and confident results in voice quality assessment. Notice that what really matters when proceeding with such assessment is to evaluate the influence of the communication system elements. However, since the quality of the system is reflected on the voice quality delivered by the system output, evaluating the voice quality is equivalent to evaluate the system performance. In this respect, the research community mostly uses the **Mean Opinion Score** (MOS) as the metric to measure the voice quality, which reflects such important subjectiveness [127].

To describe the most significant voice quality evaluation methods, it is useful to classify

them according to their nature. This classification can be done into two major categories: deterministic and statistic methods. The deterministic methods use computational algorithms and always produce the same results when repeatedly run over the same voice signals and under the same test conditions, *i.e.*, they are objective methods. The statistic methods use different people to evaluate voice samples by giving their opinions about quality, thus producing variable results from person to person and even for the same person at different times, *i.e.*, repeated results obtained by listening to the same voice sentence and test conditions, may be not constant. This kind of results reflects the subjective nature inherent to human opinions. They are known as subjective methods.

In both subjective and objective categories, there are methods where only the voice sample under evaluation is needed, *i.e.*, no reference signal is necessary for comparison, while other methods require a reference signal to be used for comparison. Reference and evaluated samples are also referred to as unprocessed and processed samples, respectively. In the particular case of the objective methods, the latter are called intrusive methods.

Subjective evaluation is characterised by obtaining opinions from users that reflect with naturalness all the characteristics of the human hearing psychophysical process. This process resorts to people specifically recruited to provide their opinions. In this context, recommendation ITU-T Rec. P.800 defines a set of methods and procedures to carry out reliable subjective evaluation [118]. In its *modus operandi*, people listen to a set of voice sentences and provide their opinion according to a given predefined scale in which each opinion corresponds to a numerical score. The average of these scores for all people constitutes the so-called MOS. While the previously mentioned objective metrics (*e.g.* SNR, RMSE) are the main metrics included in objective evaluation, such as QoS, MOS is the main metric actually used to measure the subjective quality experienced by the users. Therefore it is more relevant to refer to "Quality of Experience" rather than "Quality of Service" [128].

When the MOS scores refer to the listening quality, this is usually referred to as MOS_{LQS}[1] [129]. If the MOS scores are obtained in a conversational environment, where delays play an

[1]Listening Quality Subjective MOS

important role in the perceived intelligibility, then this is referred to as MOS_{CQS}[2].

Even though a significant number of participants should be used in subjective tests [118], a particular set of tests does not necessarily lead to exactly the same results every time it is repeated, as referred to, before. Furthermore subjective tests need to recruit and instruct people, which is expensive, time-consuming, difficult to implement and repeat and obviously not adequate for real-time quality monitoring. A possible solution to overcome this problem is to simulate subjective evaluation by using computational algorithms that model the *average* human hear. Since the repeatability of computer simulations gives invariable results under the same test conditions, these fall into the category of objective methods, as referred to, before. Such methods are also necessary to measure the accuracy of voice error concealment algorithms by enabling easy computation of the perceptual distortion [38]. Therefore objective tests, in which human intervention is not needed, are the best solutions to overcome the constraints of subjective ones [125].

Nowadays, the PESQ, defined in Rec. ITU-T P.862 and described later in section 3.4.1 of this thesis, is widely accepted as a reference objective method to compute approximate MOS scores with good accuracy [46].

Another category of voice quality evaluation methods, known as parametric methods, are not based on explicit voice signals but on a set of parameters characterising the voice communication system under evaluation[3]. The most paradigmatic parametric method is the E-Model, defined in the ITU-T Rec. G.107 [47].

Fig. 3.1 illustrates this classification of voice quality evaluation methods, according to the nature of the process used to obtain the quality scores.

[2]Conversational Quality Subjective MOS

[3]In fact the establishment of such parameters was statistically done by using voice signals to achieve to the values of such parameters.

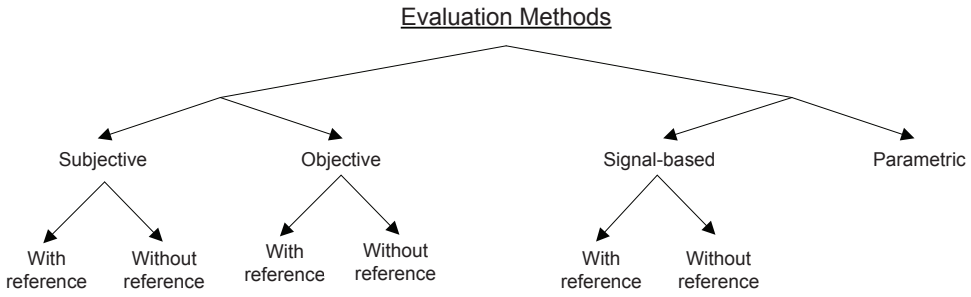


Figure 3.1 – Classification of voice quality evaluation methods

Next section describes with more detail the standard procedures defined by ITU-T.

3.3 Subjective Methods

As previously mentioned, standard subjective methods for voice quality evaluation are described in recommendation ITU-T Rec. P.800 and they are known as MOS-based methods [118]. They are considered to provide the ground-truth results because the auditory and evaluation process is carried out by human listeners [24]. This ITU-T recommendation contains the definition of the standard procedures that must be followed in order to carry out valid subjective assessment practices of voice quality in communication systems. The voice signals are presented to each listener and their individual opinions comprise the respective contributions to the MOS. The MOS value can be obtained by the following expression

$$MOS = \sum_{i}^{N} p_i \Big/ N, \tag{3.1}$$

where p_i is the score associated to the opinion given by each subject in experiment i, N is the number of experimental tests and MOS is the average of all scores. The individual opinions are not the scores but they have a direct correspondence to each score through an appropriate corresponding table – an opinion scale, as already mentioned.

Recommendation ITU-T Rec. P.800 establishes a set of mandatory procedural requirements to be met when executing the evaluation tests. Some of the most relevant requirements define the objective elements of the following aspects:

- Physical conditions of the room.
- The progress monitoring of the experiments.
- Criteria for selecting subjects. Furthermore, other aspects such as the absence of speech or hearing deficiencies, to be native speakers, or to ensure a proper male/female balancing on the number of participants must be also taken into account. The subjects should be randomly chosen from the phone user population, provided that:
  - they are not directly involved in work related to performance evaluation of telephone communications;
  - they did not participate in any subjective tests in the previous six months neither in conversation tests during the previous year;
  - they had never heard the same list of sentences of the ongoing testing.
- Procedures for recording sentences and talks, as the distance of the speaker's lips from the microphone, the recording levels to avoid overload, the use of calibration tones, annotation of the testing conditions for future comparisons with other laboratories, or the SNR.
- Hearing procedures, which include to guarantee the necessary comfort conditions to subjects and give them all the necessary instructions, such as information about the whole sequence of test steps and how to score the experienced quality.
- Characteristics of the recording and monitoring equipment, which includes the use of linear microphones, low noise and flat frequency response amplifiers.
- Gathering results and their statistical processing. Each type of test has its own type of opinion scale.

Next sections describe more specifically the variants of subjective methods and their respective opinion scales with more details.

3.3.1 Conversation-opinion tests

Conversation tests are characterised by a two-way vocal interaction between two subjects, where each one alternates the role of speaker and listener. Two opinion scales are used in

this type of quality evaluation.

The conversational quality scale contains five levels to rank the *Opinion of the connection [the subject] have just been using*, from Bad (level=1) to Excellent (level=5). Based on the individual quality levels, the quantitative scores are recorded and an average value is obtained by using expression 3.1, which corresponds to a Mean Conversation-Opinion Score represented by MOS_C[4]. Table 3.1 shows conversation opinions and respective scores to take into account.

The other opinion scale concerning conversation-opinion tests, is the difficulty scale, which allows either "Yes" or "No" as possible user opinions. These are the answers to be given by the subject at the end of each conversation to the following question; *Did you or your partner have any difficulty in talking or hearing over the connection?* This procedure leads to the Percentage Difficulty scale, represented by the symbol %D.

Besides the calculation of the average values, confidence limits should be evaluated and significance tests performed by conventional analysis-of-variance techniques. As further aid to represent the data, graphics should be plotted showing the MOS as a function of the parameter under test (*e.g.* MOS_C *versus* circuit attenuation).

Table 3.1 – Conversational opinion scale [118].

Conversation Quality	Quantitative Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

3.3.2 Listening-opinion tests

In listening tests, the subjects just listen to a set of sentences and provide their opinions about the perceived quality. The subjects do not talk, so the vocal interaction is not two-way directional as in the conversation-opinion tests. This type of tests encompasses

[4]Conversational MOS

the set of methods described below.

Absolute Category Rating (ACR) method: this method uses three opinion scales: the Listening-Quality scale, the Listening-Effort scale and the Loudness-Preference scale This method does not use any reference signal.

The Listening-quality scale, as shown in Table 3.2, contains five levels for the *quality of the speech* opinions which scores range from "Bad" (level=1) to "Excellent" (level=5). The Mean listening-quality Opinion Score is defined as MOS.

Table 3.2 – Listening-quality scale [118].

Quality of the speech	Quantitative Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

The Listening-effort scale, as shown in Table 3.3, represents the *effort required to understand the meaning of the played sentences* that led to the previous scores. These opinions range from "No meaning understood with any feasible effort" (level=1) to "Complete relaxation possible; no effort required" (level=5). According to [118] the header of the table representing the listening-effort scale is particularly important because without such a header, the descriptions are quite vulnerable to be misunderstood. The resulting average quality is represented by the symbol MOS_{LE}[5].

Table 3.3 – Listening-effort scale [118].

Effort required to understand the meanings of sentences	Quantitative Score
Complete relaxation possible; no effort required	5
Attention necessary; no appreciable effort required	4
Moderate effort required	3
Considerable effort required	2
No meaning understood with any feasible effort	1

The Loudness-preference scale, shown in Table 3.4, takes into account the perceptual level

[5]Listening-Effort MOS

of the listen speech signal. The *loudness preference opinions* range from "Much quieter than preferred" (level=1) to "Much louder than preferred" (level=5). The resulting average quality is represented by the symbol MOS_{LP}[6].

The used vocal material shall consist of simple, short sentences, meaningful and randomly chosen from non-technical literature as defined in the ITU-T Rec. P.800. A length between 2 and 3 seconds for the sentences is considered appropriate. ITU-T Rec. P.800 suggests the use of the following five sentences: "You will have to be very quiet"; "There was nothing to be seen"; "They worshipped wooden idols"; "I want a minute with the inspector" and "Did he need any money?".

Table 3.4 – Loudness-preference scale [118].

Loudness preference	**Score**
Much louder than preferred	5
Louder than preferred	4
Preferred	3
Quieter than preferred	2
Much quieter than preferred	1

Degradation Category Rating (DCR) method: this method is derived from the ACR one and improves its sensitivity to differentiate the perceived voice quality in communication systems with good quality. The DCR method uses a reference signal (signal A), ideally with a MOS score of 5, passed through the degradation system under evaluation to obtain the lower quality signal (signal B). Both voice samples are presented to listeners either as simple pairs (A-B) or in pairs with repetition (A-B-A-B). Samples A and B must be separated from 0.5 to 1 second. In a repeated procedure of pairs (A-B-A-B), time separation between two pairs should be 1 to 1.5 seconds. The *level of annoying* opinions range from "Degradation is very annoying" (level=1) to "Degradation is inaudible" (level=5). The evaluated Degradation Mean Opinion Score is represented by the symbol DMOS. Table 3.5 shows the Degradation Category scale.

[6]Loudness Preference MOS

Table 3.5 – Degradation Category Rating scale [118].

Level of annoying	Score
Degradation is inaudible	5
Degradation is audible but not annoying	4
Degradation is slightly annoying	3
Degradation is annoying	2
Degradation is very annoying	1

Comparison Category Rating (CCR) method: this method is similar to DCR: in each experiment one pair of voice samples is presented to the listener. However, in the CCR method, both the processed and unprocessed samples are randomly chosen in each test. In half of the experiments the processed sample follows the unprocessed one whereas in the remaining half this order is reversed. Also important in the CCR method is that the processed sample can be either of degraded or improved quality. The subject listeners use an opinion Comparison Scale with levels ranging from "Much Worse" (level= −3), through "About the Same" (level=0) to "Much Better" (level=3). The resulting MOS is represented by the symbol CMOS. Table 3.6 shows the opinions and respective scores in regard to the Comparison Category Rating scale.

Table 3.6 – Comparison Category Rating scale [118].

Quality of the Second Compared to the Quality of the First	Score
Much Better	3
Better	2
Slightly Better	1
About the Same	0
Slightly Worse	−1
Worse	−2
Much Worse	−3

Both DCR and CCR methods are particularly useful to evaluate the performance of telecommunication systems where the input has been corrupted by background noise. However, the main advantage of the CCR method is its ability to evaluate both degraded and improved voice quality.

Continuous Evaluation of Time Varying Speech Quality (CETVSQ) method: this method is defined in ITU-T Rec. P.880 [130]. It is particularly useful to assess

the impact of temporal fluctuations in voice quality. By taking into account both the instantaneous and the overall perceptual quality, this method is a useful tool to diagnose degradations due to either packet loss in VoIP or handover in mobile networks. It is composed by two parts:

- Formulation, at a given instant, of an opinion in a continuous scale with five levels ("Bad" to "Excellent") using a sliding device during the speech sequence.
- Formulation, at the end of the sequence, of a global opinion, on the standard ACR listening-quality scale as presented in Table 3.7.

Table 3.7 – Overall quality scale [130].

Quality of speech	Associated score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

3.4 Objective Methods

To be valid, subjective methods require both precise testing procedures and a significant number of opinions to make them representative, as previously mentioned. These requirements make such methods very onerous and thus not applicable in many real application scenarios. Therefore, a more practical and quick solution is to use objective methods to assess the subjective quality. Their repetition under the same conditions always leads to the same final quality scores and they are relatively easy to repeat [131]. These methods are intended to evaluate the subjective quality based on mathematical models of the psycho-physical system of an *average* human ear, considered as representative of the existing human diversity.

In [113] a non-intrusive hybrid method that makes use of subjective evaluation to implement an objective procedure is proposed for speech, audio and video. The underlying idea is to use several distorted samples that were previously subjectively evaluated and then use the results of this evaluation to train a Random Neuronal Network (RNN) to obtain the relation between the parameters that cause distortion and the perceived quality. That

authors claim that results given by this system have good correlation with those coming from the human perception.

Going back in the history of the perceptual voice evaluation, the Bark Spectral Distortion (BSD) is perhaps the first objective measure to take psychoacoustic factors into account [132]. It calculates an objective measure for signal degradation based on measurable properties of the auditory perception. When comparing its performance with the deterministic classic parameters (SNR, SNRseg, dB, RMSE, PSNR), it denotes a better correlation with people opinions, since it takes into account the speech loudness –a psychoacoustical parameter defined as the magnitude of auditory sensation.

Spectral domain measures are known to be better correlated with the human perception. One of their critical advantages is that they are less sensitive to signal misalignment and phase shift between the original and the distorted signals than time domain measures [9].

In [114] an evaluation methodology that combines elementary objective voice quality metrics, such as SNR and segmental SNR, energy ratio and cepstral distance among others, with a frame synchronisation mechanism, is introduced. The method was applied in a case study about the impact of Robust Header Compression (ROHC) on the voice quality achieved with real-time transmission of Global System for Mobile communications (GSM) voice over a wireless link. It permitted to measure a MOS enhancement of 0.26 for a wireless bit error probability of 10^{-3} due to the use of the ROHC strategy.

There are some research studies in the past that are worth to refer, since they led to some of the evaluation methods currently in use. In 1992 Beerends and Stemerdink developed a general method to measure the subjective quality of audio devices that uses a model of the human auditory system [133]. Such model makes transformations from the physical to the psychophysical domain performed by way of two operations: time-frequency spreading and level compression. It is known as the Perceptual Audio Quality Measure (PAQM) and it is claimed to achieve good correlation with the perceived audio quality.

Recognising that the characteristics of speech and music are different, PAQM was optimised for speech by modifying some of its procedures which led to the Perceptual Speech

Quality Measure (PSQM) method, developed by the same authors as PAQM [119]. This method is defined in ITU-T Rec. P.861 [134]. The PSQM was first developed in the KPN research labs by John G. Beerends, by 1993. It was primarily focused on identifying the quality impact of coding distortions.

The PSQM method relies on the comparison between the degraded signal and its corresponding reference signal. It transforms the speech signals into the loudness domain and by further mathematical transformations where noise disturbances (the difference between the scaled loudness of the distorted speech and the loudness of the reference speech) are calculated for all the signal frames, it estimates an average distortion. Later, in 1998 PSQM was standardised by the ITU-T as Recommendation P.861 [134]. The evaluation process takes into account distortion factors originated by noise and coding distortion introduced by lossy algorithms. Thus, it is recommended to be used only for evaluation of voice codecs. The signal bandwidth is to be comprised in the telephonic band of 300-3 400 Hz. PSQM$^+$ is an improved version of PESQ that enhances its sensitivity to loud distortions and temporal clippings.

PSQM is historically important but it does not take into account typical phenomena of telecommunication systems such as filtering, delay, jitter or channel errors. Thus, it is not suitable to evaluate voice quality along transmission systems. According to ITU-T, relative Rec. P.861 has been superseded by ITU-T Rec. P.862, herein described in section 3.4.1.

The Perceptual Analysis Measurement System (PAMS) is also an objective method that takes into consideration the human factors that affects the voice quality to evaluate the perceived speech quality of telephone networks. It was developed in the British Telecom Labs by Hollier and Rix [111, 135]. It permits to overcome the handicap of some previous methods that do not take into consideration two key network properties: linear filtering and variable bulk delay. It permits to extract and select the most representative parameters of the voice degradation. A parameter set in which the value of each parameter increases with increasing degradation is generated by means of a training procedure.

Thus, such method is flexible in choosing the parameters of interest if they are perceptually important. So its performance depends on the designer intuitive options.

Another objective method to perceptually evaluate speech quality is the QVoice, developed by ASCOM[7]. It is useful to evaluate speech quality in mobile communication systems. This method uses artificial neural networks and fuzzy logic techniques to estimate listeners' judgement of subjective quality. Its originality is to consider the Linear Predictive Coding (LPC) cepstral coefficients over a fixed duration of speech sample (5 seconds) as the perceptually significant parameters. The differences between coefficients of the reference and those of the distorted speech signals are used to train the neural network and thus estimate degradations. Then fuzzy logic is used to predict the subjective score using the estimated degradations [136].

3.4.1 Perceptual Evaluation of Speech Quality (PESQ)

The method named as "Perceptual Evaluation of Speech Quality" (PESQ) is described in the ITU-T Rec. P.862 [46]. PESQ is nowadays widely accepted as a reference objective method to compute approximate MOS scores with good accuracy [137]. It permits to achieve quality scores for narrow-band telephony codecs and end-to-end telephone systems. PESQ compares an original (unprocessed) signal, $x(t)$, with its degraded (processed) version, $y(t)$, that results from transmitting $x(t)$ through the system under evaluation. The output given by PESQ is a MOS prediction of the perceived quality, *i.e.*, an estimate of the result of real subjective testing. The key process is the transformation of both the original and degraded signals into an intermediate representation, which is analogous to the psychophysical representation of audio signals in the human auditory system. Such representation takes into account the perceptual frequency (Bark) and loudness (Sone). Then, in the Bark domain, some perceptive operations are performed taking into account loudness densities, from which disturbances are calculated. Based on these disturbances, the PESQ MOS is derived. This is commonly called the raw MOS since the respective values range from -1 to 4.5. It is often necessary to map raw MOS into another scale in order to compare the results with MOS obtained from subjective methods. The

[7] www.ascom.com

ITU-T Rec. P.862.1 provides such a mapping function, from which the so-called MOS_{LQO}[8] is obtained [138]. Fig. 3.2 shows this mapping function which results from equation (3.2). As it can be seen, it maps PESQ MOS vales ranging from −1 to 5 to MOS_{LQO} values ranging from 0 to 4.7, which are closest to the range of values obtained from real people. In fact, people do not evaluate quality with scores below zero and for a group of people it is likely that some of them do not give the maximum score, which justifies average values lower than the individual allowed maximum (5). This equation is taken from the ITU-T Rec. P.862.1 and also expresses the need to use a precision of three decimal places.

$$MOS_{LQO} = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.4945 \times rawMOS + 4.6607}}. \tag{3.2}$$

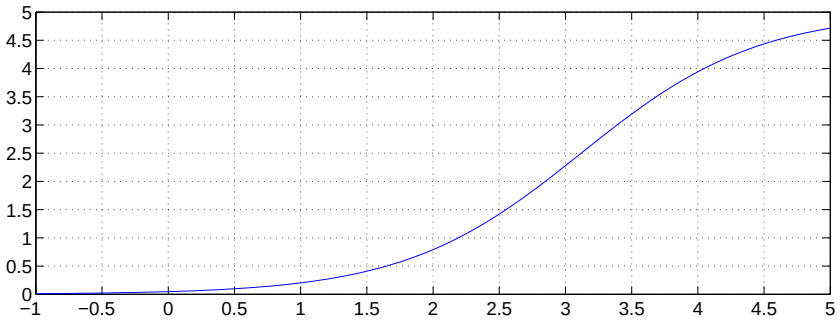


Figure 3.2 – Function that maps raw MOS to MOS_{LQO} [46].

Fig. 3.3 shows the basic operational framework of PESQ and its relationship with subjective MOS. The dotted line means that just few subjective methods require a reference signal, as in the case of CCR and DCR.

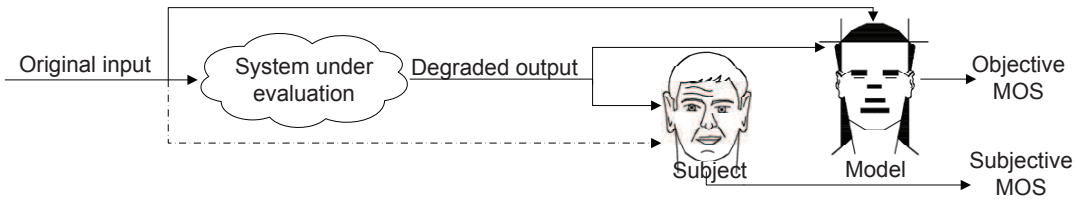


Figure 3.3 – Overview of the evaluation methodology used in PESQ [46].

[8]Listening Quality Objective MOS

PESQ takes into account the actual effects introduced by telephone systems, like filtering, delay variation and signal distortion. It does not take into account delays, so it is not useful to evaluate bidirectional systems. ITU-T Rec. P.862 defines test factors (such as variations in signal level, variations in delays, echoes, CELP codecs and hybrid codecs with rates below 4 kbps), coding technologies and applications where this method achieves accurate results. Table 3.8 (Table 1 of P.862) shows test factors, coding technologies and applications for which this method proved to have acceptable accuracy.

Table 3.8 – Factors, technologies and applications for which PESQ had demonstrated acceptable accuracy [46].

Test factors
Speech input levels to a codec
Transmission channel errors
Packet loss and packet loss concealment with CELP codecs
Bit rates if a codec has more than one bit-rate mode
Transcoding
Environmental noise at the send side
Effect of varying delay in listening only tests
Short-term time warping of audio signal
Long-term time warping of audio signal
Coding technologies
Waveform codecs, *e.g.* G.711; G.726; G.727
CELP and hybrid codecs ≥ 4 kbit/s, *e.g.* G.728, G.729, G.723.1
Other codecs: GSM-FR, GSM-HR, GSM-EFR[9], GSM-AMR[10], CDMA-EVRC[11,12],TDMA-ACELP[13,14], VSELP[15], TDMA-VSELP, TETRA
Applications
Codec evaluation
Codec selection
Live network testing using digital or analogue connection to the network
Testing of emulated and prototype networks

[9]Enhanced Full Rate
[10]Adaptive Multi-Rate
[11]Code Division Multiple Access
[12]Enhanced Variable Rate Codec
[13]Time Division Multiple Access
[14]Algebraic Code-Excited Linear Prediction
[15]Vector Sum excited Linear Prediction

Since PESQ revealed to be less sensitive to some impairment factors, such as delay, it does not provide a comprehensive assessment of the overall transmission quality. It is useful to evaluate listening-quality but not conversational-quality. Effects deriving from loss of loudness, delay, sidetone and other degradations related to two-way interaction are not taken into account by the output scores of PESQ. Thus, it is possible to obtain high PESQ scores with a poor call quality because the degraded signal may include such impairment factors, that are not relevant to PESQ. The PESQ scores have the similar characteristics and the same nature as those obtained from the listening-quality method ACR. This recommendation should be considered as referring to the opinion scale ACR_{LQO}[16] [129]. Currently PESQ is the main reference on telephony voice quality evaluation.

Wideband extension to PESQ

An extension of PESQ is the standardised ITU-T Rec. P.862.2 ("Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs") [139]. It allows ITU-T Rec. P.862 algorithm to be used in the evaluation of testing conditions, such as speech codecs with listener using wideband headphones. This recommendation is mainly intended for use with wideband audio systems (50-7 000 Hz), although it may also be useful in systems with a narrower bandwidth. Despite the fact that this recommendation is a PESQ extension, the mapping function that allows linear comparisons with MOS values produced from subjective experiments, including wideband speech, cannot be used for direct comparisons with the scores produced by baseline ITU-T Rec. P.862 or ITU-T Rec. P.862.1, due to the different experimental context [139]. Furthermore, the output mapping function from raw MOS to MOS_{LQO} used in the baseline recommendation and that used in wideband extension are different. This can be seen by comparing equation (3.2) with its counterpart equation (3.3).

$$MOS_{LQO} = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.3669 \times rawMOS + 3.8224}}. \tag{3.3}$$

In [137] a tool for quality estimation of short voice segments is proposed. It is claimed that it can be used for evaluate signals of about 5 seconds and the results are compatible with those generated by PESQ [46]. The proposed tool reduces processing delay which is useful in real time calculations when packet voice are lost.

[16] Listening Quality Objective ACR

3.4.2 Perceptual Objective Listening Quality Assessment

The Perceptual Objective Listening Quality Assessment (P.OLQA) algorithm is the next-generation voice quality testing technology for fixed, mobile and IP-based networks. It is described in the ITU-T Rec. P.863 and is the result of successive enhancements initiated with PSQM. *Grosso modo*, it can be said that the need to assess impacts of new network impairments led to PESQ; the extension of the bandwidth of PESQ from 3 100-3 400 Hz to 50-7 000 Hz led to the wideband extension of PESQ as described in the ITU-T Rec. P.862.2, and a new extension of bandwidth, among other improvements, led to the ITU-T Rec. P.863 [124]. The voice bandwidth allowed by this recommendation (named as *superwideband*) ranges from 50 to 14 000 Hz.

Two operational modes are supported: one for narrowband and the other for superwideband. Concerning the former, the listening case encompasses the use of a telephone handset while in the latter a flat headphone is assumed. The MOS scale of the former is the ACR MOS as in PESQ, while in the second it is the future oriented MOS ACR superwideband scale. MOS is now designated as $\text{MOS}_{\text{LQOsw}}$[17] and for the former it is now MOS_{LQOn}[18]. The algorithm is designed for assessing the speech quality of current and near future telephony systems that utilise a broad variety of coding, transport and speech enhancement technologies. Specifically, it allows the assessment of networks and codecs that introduce time warping; permits evaluate factors, such as packet loss concealment, listening levels, temporal and amplitude clipping or frequency response. It also introduces a broad set of new codecs such as Adaptive Multi-Rate (AMR), MPEG-2 Audio Layer III (MP3) and Advanced Audio Coding (AAC), among others. Consequently, new applications to this algorithm are now possible. Table 3.9 (Table 1 of ITU-T Rec. P.863) shows test factors, coding technologies and applications for which this method proved to have acceptable accuracy.

As in the case of the PSQM and PESQ algorithms, the approach of P.OLQA requires both the undistorted reference signal and the received signal to be scored. Therefore, it is also an intrusive method.

[17] superwideband MOS_{LQO}
[18] narrow MOS_{LQO}

Table 3.9 – Factors, technologies and applications for which P.OLQA had demonstrated acceptable accuracy [124].

Test factors
Packet loss and packet loss concealment
Acoustic noise in sending environment
Listening levels between 53 and 78 dB(A) SPL[19] in superwideband mode
Packet loss and packet loss concealment with PCM[20] type codecs
Temporal and amplitude clipping of speech
Linear distortions, including bandwidth limitations and spectral shaping ('non-flat frequency responses')
Frequency response
Coding technologies
ITU-T G.711 PLC, ITU-T G.711.1
ITU-T G.718, ITU-T G.719, ITU-T G.722, ITU-T G.722.1, ITU-T G.726
AMR-NB[21], AMR-WB (ITU-T G.722.2), AMR-WB+
PDC-FR[22], PDC-HR[23]
EVRC (ANSI/TIA[24]-127-A), EVRC-B (TIA-718-B)
Skype (SILK V3, iLBC[25], iSAC[26] and ITU-T G.729)
Speex, QCELP[27] (TIA-EIA[28]-IS-733), iLBC, CVSD (64 kbit/s, "Bluetooth")
MP3, AAC, AAC-LD[29]
Applications
Terminal testing, influence of the acoustical path and the transducer in sending and receiving direction
Bandwidth extensions
Live network testing using digital or analogue connection to the network
Testing of emulated and prototype networks
UMTS[30], CDMA, GSM, TETRA, WB-DECT[31], VoIP, POTS[32], PSTN, Video Telephony, Bluetooth
Voice Activity Detection (VAD), AGCAutomatic Gain Control (AGC)
Voice Enhancement Devices (VED), Noise Reduction (NR)
Discontinuous Transmission (DTX), Comfort Noise Insertion

[19] Sound Pressure Level
[20] Pulse Code Modulation
[21] Adaptive Multi-Rate Narrowband
[22] Personal Digital Cellular - Full Rate
[23] Personal Digital Cellular - Half Rate
[24] Telecommunications Industry Association
[25] internet Low Bitrate Codec
[26] internet Speech Audio Codec
[27] Qualcomm CELP

Like in PESQ, the key process in P.OLQA is the transformation of both the original and degraded signals into an intermediate representation, which is analogous to the psychophysical representation of audio signals in the human auditory system. In narrowband mode, the maximum ITU-T P.863 MOS_{LQO} score is 4.5 again, while in superwideband mode, it is 4.75 (MOS ACR superwideband scale is used). This maximum MOS is in line with the underlying idea that signals with larger bandwidth have better quality than those of narrow bandwidth. If a narrow bandwidth signal passes a clean system and its quality is evaluated by P.OLQA in superwideband mode, the score that might be 4.75 will be 4.5 because this algorithm takes into account bandwidth impairments and will interpret its native narrow bandwidth as it has been filtered (degraded) by the system under evaluation.

The main disadvantage of PSQM, PESQ and P.OLQA comes from the fact that, besides the degraded signal, these methods need the corresponding reference signal. Therefore, when performing speech quality evaluation in an operational communication system, these methods are intrusive in the sense that making a reference signal available at the evaluation point interferes with the normal operation of the service. Furthermore, it may be impossible to capture the reference signal, especially if it is not available at the evaluation point (*e.g.*, end-user terminal). As an alternative, non-intrusive methods are the best solutions. In the next section a non-intrusive method standardised by ITU-T is presented.

3.4.3 Single-ended method for objective speech quality assessment in narrow-band telephony applications

This method is described in the ITU-T Rec. P.563 [140] and it is used to evaluate narrowband telephony codecs and end-to-end telephone systems. Contrary to PSQM, PESQ and P.OLQA, this method does not need to be given a reference speech signal to be compared

[28]Electronic Industries Alliance
[29]AAC-Low Delay
[30]Universal Mobile Telecommunication System
[31]Wideband Digital Enhanced Cordless Telecommunications
[32]Plain Old Telephone Service

with the one under evaluation; hence this is non-intrusive, also called a single-end method. Moreover, it takes into account all distortions occurring in the PSTN, though it only measures effects of one-way transmission. Since this is a non-intrusive method, it is suited to perform speech quality evaluation in real-time applications, but its utilisation is not restricted to end-to-end conversations and it can be used to measure voice quality at any point of the transmission network. The resulting MOS is comparable to that of the perceived quality by a human listener as if he/she was listening at the same point in the network using a conventional telephone.

The key process of this algorithm is the determination of a prominent distortion class among three fixed ones. This determination permits to adjust a speech quality model, which in turn provides the appropriate quality score. Distortion classes include vocal tract analysis and unnaturalness of speech, noise analysis, interruptions and time clippings. The prominent distortion class is determined by a set of key parameters such as pitch average, SNR or LPC kurtosis among other ones. The prominent distortion class is also the class used to calculate an intermediate rough estimate of MOS. Then some additional signal features, like speech level or some specific noises, allow the computation of the final speech quality.

As in PESQ, the output given by the P.563 algorithm is an estimate of the perceived quality, given as MOS score. This MOS uses the ACR listening quality scale and must be understood as an ACR_{LQO}. As pointed out before, the P.563 algorithm takes into account the real effects of telephone systems, like transmission channel errors, delay variation or distortion. Table 3.10 (Tables 1/2/3 of Rec. P.563) shows the test factors, coding technologies and applications for which this method has acceptable accuracy. However, it is said that the disadvantage of these methods is the lower accuracy and reliability [24].

Table 3.10 – Factors, technologies and applications for which P.563 has acceptable accuracy [140].

Test factors
Characteristics of the acoustical environment
Environmental noise at the send side
Characteristics of the acoustical interface of the sending terminal
Remaining electrical and encoding characteristics of the sending terminal
Speech input levels to a codec
Transmission channel errors
Packet loss and packet loss concealment with CELP codecs
Bit rates if a codec has more than one bit-rate mode
Transcoding
Effect of varying delay on listening quality in ACR tests
Short-term time warping of speech signal
Long-term time warping of speech signal
Transmission systems including echo cancelers and noise reduction systems under single talk conditions and as they will be scored on an ACR scale
Coding technologies
Waveform codecs, *e.g.*, G.711; G.726; G.727
CELP and hybrid codecs ≥4 kbit/s, *e.g.*, G.728, G.729, G.723.1
Other codecs: GSM-FR, GSM-HR, GSM-EFR, GSM-AMR, CDMA-EVRC, TDMA-ACELP,TDMA-VSELP, TETRA
Applications
Live network monitoring using digital or analogue connection to the network
Live network end-to-end testing using digital or analogue connection to the network
Live network end-to-end testing with unknown speech sources at the far end side

3.5 A parametric method

While signal-based methods use perceptual features extracted from the speech signal to estimate quality, parametric models rely on calculations over pre-determined characteristic parameters concerning equipments in the communication chain (codecs, hybrids, handsets, etc.) and the whole technological factors involved on the communication process such as packet loss pattern, loss rate, delay and loudness. Earlier methods were developed by some telecommunication operators in the 1980 decade, when noise and attenuation were the main voice impairment factors. Nowadays, the most relevant parametric method is

the E-Model, described in the ITU-T Rec. G.107 [47], which is targeted at the modern communication networks, since it takes into account degradation factors such as delay, packet loss and coding, including noise and attenuation. It is presented as a planning tool to evaluate the combined effect of several parameters affecting the quality of 3.1 kHz bandwidth telephony conversations. This model can predict the transmitted speech quality from objective parametric values concerning physical characteristics of the subsystems involved in the acquisition, transmission, switching and delivery of voice signal. It provides an estimate of the transmission quality from the mouth of the speaker to the ear of the listener, as it would be perceived by an *average* user positioned at the "receive side". Besides its parametric nature, it may also be classified as an objective method for subjective evaluation due to the fact that input parametric values are objective and that the results predict subjective user opinions.

The output of the E-Model is a rate factor, R, which represents the overall quality of a two-way communication[33].

There is a large number of parameters inherent to transmission systems that represent degradation factors for the voice quality, influencing the overall voice quality in an end-to-end system and thus the rating factor, R.
The E-Model assumes that transmission voice impairments can be transformed into psychological impairment factors in an additive psychological scale. So, the R-factor combines all relevant transmission parameters. It is given by

$$R = R_0 - I_s - I_d - I_{e\text{-}eff} + A. \tag{3.4}$$

In expression (3.4), R_0 represents a base factor representative of the basic SNR, including noise sources such as circuit noise and room noise. I_s is named as the "Simultaneous Impairment Factor" and represents the sum of all impairments, which occur more or less simultaneously with the signal transmission. I_d is denominated as the "Delay Impairment Factor" and includes the impairments due to delay. I_{e-eff} is denominated by the "Effective Equipment Impairment Factor" and represents the impairments caused by equipment

[33]Important parameters to characterise quality encompass several path delays, which are determinant in the two-way communication.

(*e.g.*, low bitrate codecs) and random packet loss. Citing [47], A is an "advantage factor that allows for compensation of impairment factors when there are other advantages of access to the user". For example, a user of GSM technology is more tolerant to lower experienced quality than he would be when using PSTN. Thus, such user may tend to score with the same value two telephone sessions with objective difference in the quality of the received signal. It should be pointed out that quality evaluation includes the user expectation about quality, as referred to in section 3.2. Table 3.11 shows example values of advantage factor, A, according to the E-Model application scenario.

Table 3.11 – Examples of utilisation of advantage factor, A [141]

Communication system example	**Maximum value of A**
Conventional	0
Mobility by cellular networks in a building	5
Mobility in a geographical area or moving in a vehicle	10
Access to hard-to-reach locations, *e.g.* via multi-hop satellite connections	20

The R values range from 0 to 100, where 0 corresponds to the worst quality and 100 corresponds to the best quality. Based on the value of R, the ITU-T Rec. G.109 defines five categories of speech transmission quality that range respectively from "Poor" to "Best" to which correspond five categories of user satisfaction that range from "Nearly all users dissatisfied" to "Very satisfied". Table 3.12 shows such a correspondence [142].

From the R-factor, it is possible to infer an estimated conversational MOS score from the expression (3.5) [47]:

$$MOS_{CQE}{}^{34} = \begin{cases} 1, & R < 0 \\ 1 + 0.035R + R(R-60) \times (100-R) \times 7 \times 10^{-6}, & 0 < R < 100, \\ 4.5, & R > 100. \end{cases} \tag{3.5}$$

The terms R_0, I_s, I_d and I_{e-eff} encompass more specific degradation parameters that constitute the input of the E-Model algorithm. Next subsection is devoted to the description of such input parameters.

[34] Conversation-Quality Estimated MOS

Table 3.12 – Definition of categories of speech transmission quality and respective MOS_{CQE} according to [142] and expression (3.5).

R-value range	100 – 90	90 – 80	80 – 70	70 – 60	60 – 0
Quality category	Best	High	Medium	Low	Poor
User satisfaction	Very satisfied	Satisfied	Some users dissatisfied	Many users dissatisfied	Nearly all users dissatisfied
MOS_{CQE} **range**	4.5 – 4.33	4.33 – 4	4 – 3.6	3.6 – 3.1	3.1 – 2.6

By using expression (3.5) the MOS values relative to the categories referred above were added to Table 3.12.

3.5.1 Input parameters of the E-Model

In order to understand the input parameters and their nature, one should first identify the reference model connections of the E-Model algorithm. As mentioned before, the application of this algorithm assumes an end-to-end connection model between the speaker and the listener subject. Fig. 3.4 shows such a reference model. It is a generic model that constitutes the basis to understand the role of each input parameter. According to the specific application, it can be used in different scenarios, as it will be seen later in chapter 5.

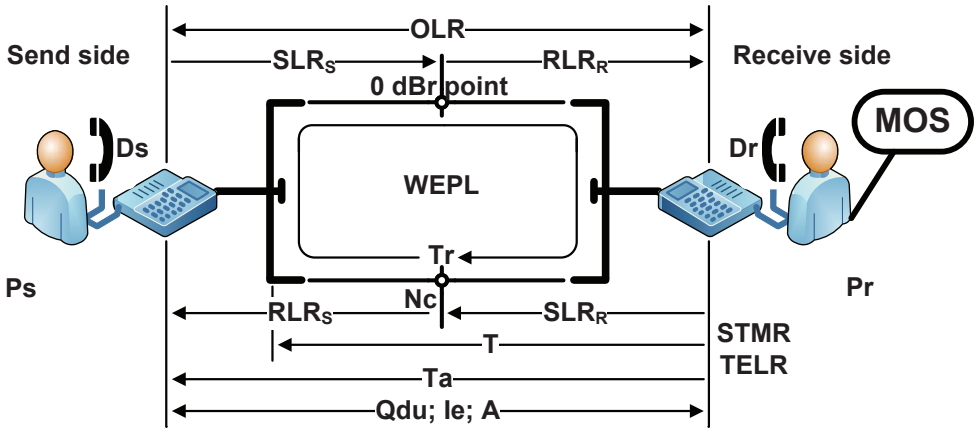


Figure 3.4 – Reference model for the E-Model [47].

Firstly, it is important to point out that this model divides the whole pathway from mouth to ear into two sections: the "send side", also referred to as side A and the "receive side", also referred to as side B. The point that makes the division is a virtual point referred to

as the "0-dBr" point, in which the signal attenuation is considered 0 dB. The location of this point depends on the scenario of interest. As an example, it can be considered as the transition point between the private and the public network [141].

From Eq. (3.4) it is possible to see that the R value is derived from the basic signal-to-noise ratio, R_0, from which degradation factor values are subtracted. R_0 represents thus the quality in the absence of impairments and is set to

$$R_0 = 94.772.^{35} \tag{3.6}$$

The input parameters can be grouped according to the impairment factors of Eq. (3.4).

Inputs that influence R_0

- **SLR**–*Send Loudness Rate* expresses the signal degradation due to attenuation (or gain) that occurs in the send side. This value can be obtained from the known characteristics of the equipments and from the transmission circuits. If this is not possible to obtain, then it is recommended to use the default value of +8 dB. It does not apply to VoIP telephones. SLR may be viewed from two points: when communication is considered to occur from the send side to the receive side and when the communication is considered to occur from the receive side to the send side. In the former case, let us designate SLR by SLR_S and in the second, by SLR_R. Such designations are represented in the Fig. 3.4.
- **RLR**–*Receive Loudness Rate* expresses the signal degradation due to attenuation (or gain) that occurs in the receive side. This value can be obtained from the known characteristics of the equipments and from the transmission circuits. If it is not possible to obtain, then it is recommended to use the default value of +2 dB. It does not apply to VoIP telephones. Similar to the SLR parameter, RLR may assume two designations: RLR_R and RLR_S (see Fig. 3.4).
- **CLR**–*Circuit Loudness Rate* expresses the signal degradation due to the attenuation that occurs in the circuits.
- **OLR**–*Overall Loudness Rate* expresses the global degradation of the signal due to the attenuation (or gain) that occurs along the whole pathway. This degradation is

[35] The way to calculate this value can be found with detail in [47].

mainly composed of three factors given by the send side telephone, by the receive side telephone and from the intermediate circuits, SLR, RLR and CLR , respectively. $OLR = SLR + RLR + CLR$.

- **Ds**–*Design factor of the send side handset* is a characteristic of the terminal equipment. The respective value must be considered only in the case of non-standard handset designs. This is an non-dimensional parameter. The default value is Ds = +3. It does not apply to VoIP telephones.
- **Dr**–*Design factor of the receive side handset.*
- **LSTR**–*Listner Side Tone*[36] *Rating*. Even it is less influenced by the mismatch impedance than the Sidetone Masking Rate (STMR)[37], this parameter reflects the impedance mismatch between analog telephones and the Private Branch Exchange (PBX) or central office to which they are connected. This value is given by the expression $LSTR = STMR + Dr$. In case it cannot be calculated, it is recommended to use the default value of $LSTR$ = +18 dB. In the case of the Digital Enhanced Cordless Telecommunications (DECT) or GSM telephones are used, the default value of $STMR$ = +13 dB is used to calculate $LSTR$. It does not apply to VoIP telephones.
- **Nc**–*Circuit noise* referred to the point 0 dBr. When transmission equipments and elements are designed according to the national and international standards, their influence on the voice quality may be neglected [47]. In cases where there are cabling sections in an analog environment with interferences, this parameter must be taken into account. In cases of digital communications, the default value of −70 *dBm0p* can be used[38]. This value can be measured at any point, given that respective adjustments are done in order to reference the value to the 0 dBr point. Roughly, such adjustments consist of either adding or subtracting the sum of all Loudness Rates of the pathway between the point where it is measured and the 0 dBr point.

[36] When the delay of the speaker echo is near zero, the reflected signal is called as *side tone.*

[37] See definition of $STMR$ in the paragraph "Inputs that influence Is" on page 59.

[38] Noise power in dBm0, measured by a psophometer. In order to take into account the subjective effect that noise causes in a user of an acoustic connection, the background noise is measured through a filter that simulates the human ear sensitivity through the different frequencies of the conventional bandwidth of the telephony (or radiophony, for the case of music). This filter is embedded into special voltmeters, measuring thus the effective (RMSE) noise value. These devices are called psophometers [143].

- **Nfo**–*Noise floor* at the 0 dBr point[39]. *Nfo* can typically be obtained from the expression $Nfo = Nfor + RLR$, where
  - **Nfor**–*Noise floor at the receive side* represents the environmental noise at the receive side. *Nfor* is usually set to *Nfor*= –64 dBmp.
- **Ps**–*Room Noise at send side* expresses the room noise in the surrounding space around the telephone at the send side. It influences the signal-to-noise ratio that is perceived in the receive side. Values comprised between 30 and 50 dB(A) are considered as normal for office environments so that degradations due to this type of noise are considered not important. In these cases, the default value of +35 dB(A) may be used. However in industrial environments an average value obtained from diverse measurements must be used.
- **Pr**–*Room Noise at receive side* expresses the room noise in the surrounding space around the telephone at the receive side. It influences the perceived quality via the sidetone path. As in the case of the Room Noise at send side, the default value of 35 dB(A) may be used in the case of office environments.

Inputs that influence Is

- **STMR**–*Sidetone Masking Rate* is a characteristic of the analog terminal equipment. It is related to the degree of the matching between the balance impedance of the telephone circuit and the input impedance of the PBX line interface in conjunction with the ports that constitute the interface with 2-wire facilities. The consequence is that it may cause a non-optimum sidetone. The default value of +15 dB may be used in Europe unless there is a high probability to have a significant impedance mismatch. The default value to North America is +18 dB and +13 dB for the technologies DECT, GSM full rate, GSM half-rate or GSM enhanced full rate. It does not apply to VoIP telephones.
- **qdu**–*quantisation distortion unit* is a distortion measure due to a complete process of encoding from analog to digital (A/D) and again from digital to analog (D/A) according to the Recommendation G.711. Each pair A/D–D/A represents one *qdu*

[39] **Nfo** not represented in Fig. 3.4.

unit when using A-law or μ-Law. For coding laws other than A-law or μ-law, qdu must be replaced by equipment impairment factor (Ie), which values are fixed in the ITU-T Rec. G.108 [141]. When the elements that affect the coding are part of the connection, such as echo canceling devices, a value of qdu = 0.7 must be used for each element. If the entire end-to-end connection is digital, a quantisation distortion unit (qdu) value of qdu = 1 must be considered in the calculations, regardless whether the codec is located inside the digital telephone or in a card to adapt an analog telephone to the digital network.

- **TELR**–*Talker Echo Loudness Rating* expresses the echo signal level generated by the speaker, *i.e.*, phenomenon by which the talker hears his own voice. This value must be referred to the receive side. The degree of annoyance depends both from the delay and the difference of level between the original voice and the received echoed voice. This difference is expressed by the TELR according to the expression: $TELR = SLR_s + EL + RLR_s$ where $\mathrm{SLR_s}$ and $\mathrm{RLR_s}$ concern the talker's telephone set and EL concerns the echo loss of the echo path. The default value is $TELR$ = +65 dB.
  - **EL**–*Echo Loss* represents the sum of the losses occurred along the bidirectional echo pathway. Losses must be included twice in the calculations. The extension of this pathway depends if the configuration is 2-wire or 4-wire.

Inputs that influence Id

- **T**–*Mean one-way delay of the echo path* expresses the mean one-way delay the signal takes to trip within the echo path. It corresponds to the half of the delay recurred by the signal from the receive side to the hybrid located in the send side and return back to the telephone located in the receive side, in the case the forward and return paths are the same. Otherwise, the average of these delays must be considered. This value can be obtained by measuring it in the real environment. Otherwise the default value of T = 0 ms may be used.
- **Tr**–*Round trip delay in a 4-wire loop* concerns the round trip delay in a closed 4-wire loop (between two hybrids in the case of analog telephones). This value can

also be obtained by actual measurement. Otherwise the default value of $Tr = 0$ ms may be used.

- **Ta**–*Absolute delay in echo free connections* expresses the end-to-end delay between the send side and the receive side telephones. This value can also be obtained through measurement. Otherwise the default value of $Ta = 0$ ms may be used.
- **WEPL**–*Weighted Echo Path Loss* expresses the signal losses that occur in the round trip path (corresponds to the parameter Tr). This value can be obtained by summing all the losses in the echo path. In the case the terminal equipment is a digital telephone, the calculation of this value needs a new parameter; that is, the
  - **TCLw**–*weighted Terminal Coupling Loss*, which is a characteristic of the used telephone.

  The default WEPL value is $WEPL = 110$ dB.

Inputs that influence *Ie_eff*

- **Ie**–*Equipment Impairment Factor* is a factor to take into account in the case the used codecs are not those referred in the ITU-T Rec. G.711. In such case, the degradation must be expressed in terms of Ie instead of *qdu*. ITU-T Rec. G.108 provides the Ie values according to the codec in use. For example if the codec is G.729, Ie=10; if it is the GSM full-rate, Ie=20; if it is the G.723.1 (5.3 kbps), Ie=19. The bigger the Ie value, the more the signal degradation is.

  In the case of codecs operating in the presence of random packet losses, the Packet-loss Robustness Factor, *Bpl*, is defined as a codec-specific value as well as the Packet-loss Probability, *Ppl*, to represent the probability of packet losses. Thus, the packet-loss dependent Effective Equipment Impairment Factor, *Ie-eff*, is derived using the codec-specific value, *Ie*, for the Equipment Impairment Factor at zero packet-loss, as well as the *Bpl* and *Ppl* values. *Bpl*, *Ppl* (and also *Ie*) are tabulated in Appendix I of the ITU-T Rec. G.113 [144].

According to ITU-T Rec. G.107, *Ie-eff* can be calculated as follows:

$$Ie\text{-}eff = Ie + (95 - Ie)\frac{Ppl}{\frac{Ppl}{BurstR} + Bpl}, \tag{3.7}$$

where $BurstR$ stands for Burst Ratio, which represents how bursty the packet loss is, that is, how long a sequence of lost packets is. It is defined as

$$BurstR = \frac{Average\ length\ of\ observed\ burst\ in\ an\ arrival\ sequence}{Average\ burst\ length\ expected\ for\ the\ network\ under\ \text{“}pure\ random\text{”}\ losses}. \tag{3.8}$$

By “pure random” it is meant that the loss probability of each packet does not depend from the previous one; that is, loss events are statistically independent. In such a case, both numerator and denominator have the same value and $BurstR$ becomes $BurstR = 1$. When packet loss is bursty, $BurstR > 1$. It means that loss events are statistically dependent, that is, the occurrence of a packet loss depends whether the previous one was lost or not. In other words, once a loss has occurred, it is more likely that another one will occur than if they occur “purely random”. In practice it is found that packet losses are not “pure random”.

Example

As an example, consider a 18-packet stream represented by 000100110000111000 where the 12 zeros represent packet arrivals and the 6 ones represent packet losses. Despite this stream is not big enough to be statistically representative, it can serve to illustrate the concept. In this case $Ppl = 6/18 = 0.33$. Considering that this packet loss distribution corresponds to a 2-state Markov model[40], the respective transition probability between the non-loss state and the loss state, p, can be calculated as the ratio between the number of occurred transitions $0 \to 1$, and the number of possibilities this could generate, which coincide with the number of zeros. It is possible to infer, by inspection, that $p = 3/12 = 0.25$. Similarly, the transition probability between the loss state and the non-loss state, q, becomes $q = 3/6 = 0.5$. Since, for this distribution, $BurstR = 1/(p + q)$ (see [47]), $BurstR = 1.33$. Considering that the G.729 codec is used, $Ie = 10$ and $Bpl = 19$, so, by

[40]Not “pure random”, but losses statistically dependent.

means of Eq. (3.7), $Ie\text{-}eff$ becomes

$$\begin{aligned} Ie\text{-}eff &= 10 + (95 - 10)\frac{0.33}{\frac{0.33}{1.33} + 19}; \\ Ie\text{-}eff &= 11.45. \end{aligned} \tag{3.9}$$

In the absence of other impairments factors, except these packet losses, the rating factor becomes, by means of Eq. (3.4) and Eq. (3.6),

$$\begin{aligned} R &= 94.772 - 0 - 0 - 11.45; \\ R &= 83.321. \end{aligned} \tag{3.10}$$

By using Eq. (3.5) it is possible to derive the MOS value:

$$\begin{aligned} MOS &= 1 + 0.035 \times 83.321 + 83.321 \times (83.321 - 60) \times (100 - 83.321) \times 7 \times 10^{-6}; \\ MOS &= 4.143. \end{aligned} \tag{3.11}$$

By using the ACR scale[41] this MOS value means that an *average* user would give the opinion "Good" if prompted to evaluate such a voice quality. Notice this opinion is the second best opinion in the five of that scale.

Using the R factor and the categories defined in Table 3.12, such a connection category would fall into the "High" category which is also the second best category of five of that scale. This evaluation is in line with that using the ACR scale, which corroborates the coherence of both metrics. Finally notice that from Eq. (3.7) it is possible to see that in case of no packet losses, *Ie-eff* becomes Ie as given by ITU-T Rec. G.113.

If delay impairments are not considered, the I_d factor is not taken into account, and by means of ITU-T G.107 Annex B expressions, MOS_{CQE} is referred to as MOS_{LQE}[42] [47].

As described above, the E-Model algorithm can be useful to evaluate voice communications under production, given that they can be characterised either by the equipment parameters that constitute the communication chain, such as loudness ratings, and/or network parameters such as losses and delays. The more the number of specific input

[41] See section 3.3.2, page 38.
[42] Listen-Quality Estimated MOS

parameter values are known, the less the number of default values that have to be used and consequently the better accuracy that can be achieved. Table 3.13 shows the recommended default values [141].

Table 3.13 – Default, minimum and maximum values of the E-Model input parameters [141]

Input parameters	**Default value**	**Minimum value**	**Maximum value**
SLR	+8 dB	0 dB	+18 dB
RLR	+2 dB	−5 dB	+14 dB
CLR			
OLR			
Ds	+3	−3	+3
Dr	+3	−3	+3
LSTR	+18 dB	+13 dB	+23 dB
Nc	−70 dBm0p	−80 dBm0p	−40 dBm0p
Nfo	(Nfor = −64 dBmp)		
Ps	+35 dB	+30 dB	+50 dB
Pr	+35 dB	+30 dB	+50 dB
STMR	+15 dB (Europe)	+10 dB	+20 dB
qdu	1	1	14
TELR	+65 dB	+5 dB	+65 dB
T	0 ms	0 ms	500 ms
Tr	0 ms	0 ms	1000 ms
Ta	0 ms	0 ms	1000 ms
WEPL	+110 dB	+5 dB	+110 dB

3.6 Discussion

The most relevant methods to evaluate telephony voice quality were described on the previous sections. These methods cover the most important scenarios currently used in voice communications since they encompass both classical objective and subjective metrics and scenarios comprising both analog and digital contexts, user listening and conversation roles, narrow and wideband signals, using reference signal and single-ended methods, making use of voice samples or *simply* using system characteristics as input parameters.

The traditional objective metrics such as RMSE and SNR, despite not considered as having good correlation with the perceived quality, they are still of considerable importance

since they permit to compare systems with oldest reference evaluations and, since they are ease to use, they can be a reasonable approach when the more complex subjective evaluation is not possible to carry out. They can also be a valuable complement when used in conjunction with subjective scores.

Among the methods used to provide measures of subjective evaluation, PESQ is indubitably a reference to take into account due to its widespread use and also because it provides a universal and reliable metric for comparing results. Furthermore, even if it is not directly used, it is recommended to be used as reference when developing other voice evaluation systems. The drawback is its intrusive nature and limitation to narrowband signals. However it can be used to calibrate developed non-intrusive methods. To accomplish with the assessment needs of future wideband communication demands, P.OLQA is a promising option, since it allows to evaluate voice quality within bandwidth ranges from 50 to 14 000 Hz.

The E-Model can also be used as an evaluation method for voice quality, with three important advantages:

- It is a non-intrusive method;
- Its evaluation results include the effects of temporal impairments, such as delay, which grants it the ability to evaluate conversational sessions;
- There are no claimed intellectual property rights, so far, so it can be freely used as a base to derive suitable methods for commercial purposes.

Furthermore, successive enhancements added since the 1998 version till the 2005 version added the ability to take into account impairments due to random packet-loss for different codecs. In fact, the addition of packet-loss dependent factors such as the Packet-loss Robustness factor, *Bpl*, the Packet-loss Probability factor *Ppl*, and the Burst Ratio factor, *BurstR*, permitted to define the new Effective Equipment Impairment Factor parameter, *Ie-eff*, which measures the impairments due to random packet losses and makes the E-Model suitable to assess the quality of packet switching telephony systems as is the case of VoIP [47].

The use of these convergence of methods permitted to drive the study and subsequently derive the practical evaluation method that is described in chapter 5.

3.7 Conclusions

In this chapter the main standard methods for evaluating the quality of voice communications were presented. Concepts such as MOS and opinion scales were introduced. The basics to understand the voice quality assessment problem were addressed and practices to accurately carry out tests to measure the voice quality by means of the most adequate metric – the MOS, were described. Subjective, objective and parametric methods were presented, providing useful background knowledge for further research and development of a real experimental evaluation setup.

4 Linear Interpolation Algorithms for Signal Reconstruction

This chapter presents an investigation on linear interpolation algorithms for application in voice signal reconstruction. It starts by providing the necessary background of algebraic fundamentals that are most relevant to understand the nature of the algorithms and their mathematical formulation. Then two specific algorithms are described, namely, a maximum and a minimum dimension. The important concepts associated with signal reconstruction algorithms are presented and the respective *modus operandi* are also described. In the last section simulation results obtained from these algorithms are discussed in the VoIP context.

4.1 Algebraic Fundamentals

This section presents the most relevant concepts of linear algebra in regard to the voice reconstruction methods described in the next section. The mathematical definitions and relationships are explained with particular emphasis on applications in signal reconstruction problems. Some of these concepts are explained with further detail in [145].

Let us define $\mathbb{C}$, $\mathbb{R}$ and $\mathbb{Z}$ as the sets of complex, real and integer numbers respectively, and $\mathbb{C}^N$, $\mathbb{R}^N$ and $\mathbb{Z}^N$ as complex, real and integer N-dimensional spaces. An element of any of these sets is called a vector of dimension N. Let us consider f a continuous

function. An indexed sequence $x[n]$ given by

$$x[n] = f(nT),\ n \in \mathbb{Z}, \quad T \in \mathbb{R} \tag{4.1}$$

is defined as a sampled version of f.

A complex sequence of length N is represented by the column vector $x \in \mathbb{C}^N$ with components $[x_0, x_1, \cdots, x_{N-1}]^T$, where x^T is the transpose of x. In digital signal processing, such vector components are known as signal samples.

The solution of many signal processing problems is often found by solving a set of linear equations, *i.e.*, a system of n equations and n variables $x_1, x_2, \cdots x_n$ defined as

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases}, \tag{4.2}$$

where elements a_{ij}, $b_i \in \mathbb{R}$.

The Eq. 4.2 can be written in either matricial form,

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \ddots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{bmatrix} \tag{4.3}$$

or in compact algebraic form:

$$Ax = b. \tag{4.4}$$

A is known as the system matrix and it is said to be a $n \times n$ matrix or simply $A_{n\times n}$.

Let us consider a more generic matrix A, which number of rows is m and number of columns is n. That is, A is a $A_{m\times n}$ matrix. Let the complex number $z^* = a - b\mathbf{i}$ be the conjugate of $z = a + b\mathbf{i}$, where $\mathbf{i}$ is the imaginary unit. The conjugate transpose of the $m \times n$ matrix A is the $n \times m$ matrix $\left(A^T\right)^*$ obtained from A by taking the transpose and then

the complex conjugate of each transposed element, a_{ji}. For real matrices, $(A^T)^* = A^T$. A is normal if $A^T A = AA^T$. Any matrix A, either real or complex, is said to be hermitian if $(A^T)^* = A$. If $A^T = A$, then it is called a symmetric matrix.

A $n \times n$ matrix A is said to be diagonal if the entries that are outside the main diagonal are zero. If all the elements of that diagonal are set to 1, it is called an identity matrix, which is herein represented by I.

Denoting by I_n a $n \times n$ identity matrix, any $n \times n$ square matrix A is invertible or non-singular when there is a matrix B that satisfies the condition $AB = BA = I_n$. Matrix B is called the inverse of A, and is denoted by A^{-1}.

If A is invertible, then $A^{-1}Ax = A^{-1}b$ and the system equations $Ax = b$ expressed in 4.4 has a unique solution given by

$$x = A^{-1}b. \tag{4.5}$$

A $n \times n$ complex matrix A that satisfies the condition $(A^T)^* A = A(A^T)^* = I_n$, (or $x = A^{-1}b = (A^T)^* b$) is called a unitary matrix.

Considering a $m \times n$ matrix A and the index sets $\alpha = \{i_1, i_2, \cdots, i_p\}$ and $\beta = \{j_1, j_2, \cdots, j_q\}$, with $p < m$ and $q < n$, a sub-matrix of A, denoted by $A(\alpha, \beta)$, is obtained by taking those rows and columns of A that are indexed by α and β, respectively. For example

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} (\{1,3\}, \{1,2,3\}) = \begin{bmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \end{bmatrix}. \tag{4.6}$$

If $\alpha = \beta$, the resulting sub-matrix is called a principal sub-matrix of A.

An eigenvector v of a square matrix A is a non-zero column vector that satisfies the following condition:

$$Av = \lambda v, \tag{4.7}$$

for a scalar λ, which is said to be an eigenvalue of A corresponding to the eigenvector v. In other words, when A is multiplied by v, the result is the same as a scalar λ multiplied by v. Since it is much easier to multiply a scalar by a vector than a matrix by a vector,

the result of Eq. 4.7 is useful to use when computational requirements are an issue, given that eigenvalues are known *a priori*.

Considering that computing eigenvalues is more demanding than multiplying a matrix by a vector, the use of eigenvalues does not represent, apparently, a computation gain. However, in the cases where these values may previously be computed and the values of Av are repeatedly needed, an economy of scale may be achieved by computing λv and global computation requirements decrease, as a whole. The advantage of using eigenvalues can go further in the case where the matrix A is circulant (see page 72), since its structure gives the computation of its eigenvalues so simple as calculating the Discrete Fourier Transform (DFT) elements of its first row [146].

The spectrum of A is defined as the set of its eigenvalues, while the spectral radius of A, denoted by $\rho(A)$, is the supremum[1] among the absolute values of its spectrum elements. Since the number of eigenvalues is finite, the supremum can be replaced with the maximum. That is

$$\rho(A) = \max_i |\lambda_i|. \tag{4.8}$$

The eigenvalues of a matrix A have an important role on the solution of the system of equations. If $\rho(A) < 1$, then the inverse of $(I - A)$ exists and Eq. 4.4 has a possible solution. This solution can be obtained by a direct calculation method as given in Eq. 4.5 or by an iterative method.

A vector norm can be thought of as the length or magnitude of that vector x. Several types of norms are defined [145]. The most familiar norm is the Euclidean l_2-norm, defined as

$$\|x\|_2 = \left(\sum_{i=1}^{N} x_i^2 \right)^{1/2}. \tag{4.9}$$

Other norms, such as the l_∞-norm and l_1-norm are also relevant:

$$\|x\|_\infty = \max_{i \leq N} |x_i|, \tag{4.10}$$

[1]The supremum of a set S is v if and only if: i) v is an upper bound for S and ii) no real number smaller than v is an upper bound for S [145].

$$\|x\|_1 = \sum_{i=1}^{N} |x_i|. \tag{4.11}$$

The matrix norm subordinate to a vector norm is defined as

$$\|A\| = \sup\left\{\|Au\| : u \in \mathbb{R}^N,\ \|u\| = 1\right\}. \tag{4.12}$$

Conditioning of a problem is another important concept, informally used to indicate how sensitive the solution is to small changes in the input data. A problem is said to be ill-conditioned if small changes in the input data produce large variations in the solution, whereas the solution of a well-conditioned problem is less sensitive to variations in the input data.

For certain types of problems, a condition number can be defined as follows. Concerning the system equation defined in Eq. 4.4, a perturbation on b, $\tilde{b}$, will produce a corresponding perturbation on x, $\tilde{x}$. Thus Eq. 4.4 can be written as

$$A\tilde{x} = \tilde{b}, \tag{4.13}$$

where $\tilde{x}$ stands for the perturbation on x caused by the perturbation $\tilde{b}$ on b. The relation between relative perturbations is given by

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \|A\|.\|A^{-1}\|.\frac{\|b - \tilde{b}\|}{\|b\|} \tag{4.14}$$

which permits to define the condition number as expressed by [145]

$$k(A) = \|A\|.\|A^{-1}\|. \tag{4.15}$$

Thus, if the condition number, $k(A)$, is large, even a small error in b may cause a large error in x. If the condition number is small, then the error in x will not be much higher than the error in b. The condition number is a property of the problem obtained from matrix A, which leads to well-conditioned problems whenever its value is close to unity. In the case where A is a normal matrix, the condition number assumes the form

$$k(A) = \left|\frac{\lambda_{max}(A)}{\lambda_{min}(A)}\right|. \tag{4.16}$$

The eigenvalues of the system matrix, A, and the relation between them play an important role in the problem conditioning. In this context, special attention should be paid to the spectral radius. As it will be explained in section 4.3, a spectral radius near or greater than 1 leads to an ill-conditioned problem, whereas a spectral radius between 0 and 1 leads to a well-conditioned problem.

Another important property is idempotence by which an operation can be repeated over the same data without changing the result. In the algebra context, a $n \times n$ matrix A is said to be idempotent if $A^2 = A$.

In a Toeplitz matrix A, each of its elements satisfies $a_{ij} = a_{i-j}$, which is equivalent to $a_{ij} = a_{i-1,j-1}$, thus, each descending diagonal from left to right is constant, as shown below.

$$\begin{bmatrix} a_0 & a_{-1} & a_{-2} & \cdots & a_{-(N-1)} \\ a_1 & a_0 & a_{-1} & \cdots & a_{-(N-2)} \\ a_2 & a_1 & a_0 & \cdots & \ddots \\ \vdots & \vdots & \ddots & \ddots & a_{-1} \\ a_{N-1} & a_{N-2} & \cdots & a_1 & a_0 \end{bmatrix}. \tag{4.17}$$

In the case of a complex matrix where a_{-k} is the conjugate of a_k, then it is called an hermitian Toeplitz whereas if the matrix is real, then it is a symmetric Toeplitz. In the system Eq. 4.4, if A is a $m \times n$ Toeplitz matrix, then the system has only $m+n-1$ degrees of freedom, rather than $m \times n$. A circulant matrix is a special kind of Toeplitz matrix where each row vector is rotated one element to the right relative to the preceding row vector.

A matrix A is positive definite if the associated quadratic form is positive, *i.e.*, if $x^H Ax > 0$, $\forall x \neq 0$. If A is positive definite and symmetric, then all of its eigenvalues λ_i are real and positive [145]. Every positive definite matrix is invertible and its inverse is also positive definite [146]. A matrix A is non-negative definite if the associated quadratic form is non-negative, that is, if $x^H Ax \geq 0$, $\forall x \neq 0$.

Direct computation methods

There are several possible methods to find the solution of Eq. 4.4, which may be classified in either direct, iterative or semi-iterative methods. Concerning the direct methods, the solution can be found by left-multiplying both members of the Eq. 4.4 by A^{-1}, if it exists, resulting in Eq. 4.5. There are several possible approaches going from Gauss-Jordan elimination to factorisation methods such as Lower Upper (LU) decomposition. Some special structures of A and the characterisation of the matrix A according to the explained classification can lead to simple solutions. As an example, Eq. 4.4 has a trivial solution when matrix A is diagonal. In this case, the solution is given by

$$x = \begin{bmatrix} b_1/a_{11} \\ b_2/a_{22} \\ \vdots \\ b_n/a_{nn} \end{bmatrix}. \tag{4.18}$$

If $a_{ii} = 0$ and $b_i = 0$, for any i, then x_i may be any real number. If $a_{ii} = 0$ and $b_i \neq 0$ there is no solution for the system. If the entries below or above the main diagonal of a $n \times n$ matrix A are zero, then A is either a lower (L) or upper (U) triangular matrix, respectively, as follows:

$$L = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{21} & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix} \quad U = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix}. \tag{4.19}$$

If all entries of the main diagonal of a triangular matrix are zero, then such matrix is called either strictly upper or strictly lower triangular, respectively.
Assuming a lower triangular matrix A and $a_{ii} \neq 0$, $\forall i$, equations from Eq. 4.4 become

$$\begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{bmatrix} \tag{4.20}$$

and then obtaining x_1 from the first equation becomes trivial.

LU decomposition is a simplified method to solve a system of equations where matrix A can be defined as the product between a lower triangular matrix and an upper triangular matrix, *i.e.*, $A = LU$. Thus, the linear set of equations to solve becomes $Ax = (LU)x = L(Ux) = b$ and the solution can be found by first solving it for vector y such that $Ly = b$ and then solving for x, using $Ux = y$. The advantage of breaking up one linear set of equations into two successive ones is that the solution of a triangular set of equations is quite trivial [147].

A particular case of LU decomposition is the Cholesky decomposition where decomposition of matrix A is given by the product of a lower triangular matrix with its conjugate transpose, *i.e.*, $A = L(L^T)^*$ and L is a lower triangular matrix with all diagonal elements positive. When applicable, the Cholesky decomposition is about twice as fast as other methods used for solving systems of linear equations [148]. Note that this method requires that A is real, symmetric and positive-definite.

Iterative methods

Direct methods to resolve the type of equations such as Eq. 4.5 ideally produce an exact solution within the machine accuracy. However, when the system order is high, with thousands of equations, the computational effort may be critical either in terms of execution time or other resources like memory. In this case, iterative methods might be the answer to overcome such constraints. In their *modus operandi*, iterative methods produce a sequence of vectors that converge to the final solution vector as the computational process evolves. The process halts when either some pre-defined number of iterations is reached or an acceptable level of accuracy is obtained at any possible iteration. In high dimensional systems, if precision is not a strong requirement, it is possible to find the approximate solution with just a few iterations. Particularly, in sparse systems, where the number of zero entries in the iteration matrix is high, iterative methods prove to be very efficient in the sense that only a small number of computations are necessary.

There are several specific iterative methods particularly suited to solve systems of the

form of Eq. 4.4. Among them, **Jacobi** and **Gauss-Seidel** methods are paradigmatic. The **Jacobi** method follows from the individual analysis of each of the n system equations as defined in Eq. 4.4. If the following expression holds for the i^{th} equation,

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \tag{4.21}$$

then x_i can be solved by assuming that other entries do not vary, *i.e.*,

$$x_i = \left(b_i - \sum_{j=1}^{n} a_{ij} x_j\right) \Big/ a_{ii}, \quad j \neq i, \tag{4.22}$$

which suggests an iterative resolution for the i^{th} equation, as given by

$$x_i^{(k)} = \left(b_i - \sum_{j=1}^{n} a_{ij} x_j^{(k-1)}\right) \Big/ a_{ii}, \quad j \neq i. \tag{4.23}$$

The **Gauss-Seidel** method can be seen as an enhancement of Jacobi method in which updated values of x_i on the left-hand side of Eq. 4.23 are used as soon as they become available even in the same iteration. That is, instead of using x_j from iteration $k-1$ in iteration k, the value of x_j from previous equation is used in the same iteration when available. The first equation is the only exception. As an example, for the first two equations, we have

$$\begin{aligned} a_{11} x_1^{(1)} &= b_1 - a_{12} x_2^{(0)} - a_{13} x_3^{(0)} \cdots - a_{1n} x_n^{(0)} \\ a_{22} x_2^{(1)} &= b_2 - a_{21} x_1^{(1)} - a_{23} x_3^{(0)} \cdots - a_{2n} x_n^{(0)}, \end{aligned} \tag{4.24}$$

$$\vdots \tag{4.25}$$

which leads to

$$x_i^{(k)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)}\right) \Big/ a_{ii}. \tag{4.26}$$

From iteration $k-1$ to iteration k, the result converges to the final solution an amount given by $(\delta x)^{(k)} = x^{(k)} - x^{(k-1)}$. Thus, the solution at iteration k is given by $x^{(k)} = x^{(k-1)} + (\delta x)^{(k)}$. However, at iteration $k-1$ the result $x^{(k-1)}$ differs from the final solution the amount of $(\Delta x)^{(k)} = x - x^{(k-1)}$, which is greater than $(\delta x)^{(k)}$, if the solution converges. Since

$(\delta x)^{(k)} < (\Delta x)^{(k)}$, one may speed up the convergence rate by using the over-relaxation form of Eq. 4.27 instead of simply computing $x^{(k)}$ as $x^{(k)} = x^{(k-1)} + (\delta x)^{(k)}$. That is,

$$x^{(k)} = x^{(k-1)} + (\omega x)^{(k)}, \quad \omega > 1, \tag{4.27}$$

where w is called the relaxation factor. Typically, this value is constant for all k and $1<w<2$. The iterative process expressed in Eq. 4.27 is called Successive Over-Relaxation, commonly abbreviated by SOR.

The basic concepts of linear algebra presented above are used in problems of voice signal reconstruction dealing with missing samples, such as those described in sections 4.2 and 4.3 [63, 149–152]. These problems can be defined as linear system equations in which interpolation algorithms play an important role in finding their solutions. Next section concerns with the description of two linear interpolation algorithms.

4.2 Two linear interpolation algorithms

This section describes two reconstruction algorithms capable of computing accurate estimates of missing samples in voice signals due to packet loss or transmission errors. These reconstruction algorithms are particularly suitable to be implemented in receivers as recovery methods of lost samples to enhance the voice QoE delivered to users. In order to understand such algorithms, some mathematical concepts are first presented. They are useful to apply on laboratorial experiences that modulate the real situation in which voice samples are lost as well as in a real system.

Based on Eq. 4.1, let us define a N-dimension signal vector with components $x_1, x_2, \cdots, x_N$, to modulate a voice packet with payload of length N. Let us also define the Fourier matrix, F, a unitary $N \times N$ matrix with components F_{mk} given by

$$F_{mk} = \frac{1}{\sqrt{N}} e^{-\mathbf{i}\frac{2\pi}{N}mk}, \tag{4.28}$$

where $\mathbf{i}$ is the imaginary unit. Therefore, the DFT of x, here represented by $\hat{x}$, is the sequence $\hat{x} = Fx$.

In this context, two relevant linear operations are defined in $\mathbb{C}^N$: sampling and band-limiting [63]. The herein defined sampling concept goes beyond the usual concept by

which an analog signal is converted into a digital signal by satisfying the Nyquist theorem, for example. It refers to the operation in which some of the samples of a digital signal are set to zero. The resulting signal (called the observed signal) is hence a sampled version of the original. In this case, sampling can be seen as a mapping function that converts a sequence of samples (the original digital signal) into another one in which some of the original sample values were set to zero. This operation is intended and useful to modulate the real situation in which transmitted voice samples are corrupted or missing in known positions due to packet losses in erasure channels. Mathematically, it can be represented by multiplying the previously defined N-dimension signal vector (the original signal) by a diagonal matrix, D, whose elements are comprised of zeros and ones. The resulting signal is called the observed signal and corresponds to a corrupted version of the original one. For this reason D is called the sampling matrix and its diagonal is the sampling set associated with this sampling operation [63]. D models the real distribution of the missing samples in a practical case, so the determination of such matrix must be achieved by the real system that identifies the locations of the original signal in which samples are corrupted or missing.

Considering s the number of nonzero entries in the sampling set, then s/N defines the density of sampling. Here it is assumed that $s < N$ and D is not the identity matrix, I [63]. It represents the time domain known information.

Band-limiting can also be viewed as a sampling operation, in which the signal samples set to zero are in the Fourier domain, *i.e.*, signal frequency components. In fact, by multiplying a diagonal matrix Γ by Fx, the resulting matrix ΓFx has zeros in those spectral components of x that correspond to the zeros of Γ. It represents the spectral known information. Then left-multiplying F^{-1} by ΓFx returns the signal into the time domain, resulting in a filtering operation. Therefore, such band-limiting operation can be defined by a linear operator characterized by a matrix B defined as $B = F^{-1}\Gamma F$. Similarly to the matrix D, Γ is a sampling matrix different from the identity, I. The bandwidth of the signal $y = Bx$ is defined as q/N, where q is the number of nonzero entries in Γ.

Returning to the usual concept of sampling, the Nyquist sampling frequency is denoted

as f_s while f_{os} is an oversampling frequency: that is, $f_{os}>f_s$. In this case, an oversampling factor r is defined as $r = f_s/f_{os}$. Such oversampling factor is also given by $r = q/N$ and if $r<1$ then there is redundancy in the signal. Considering N the total number of samples of a voice signal and n the number of corrupted samples, then the following condition holds: $n < N$.

The error geometry is defined as the pattern of missing samples within the whole sequence of samples. Depending on the relative position between missing samples, three geometries are addressed in this thesis:

- **Interleaved geometry**, where the missing samples are equidistant and multiple of an integer $l \geq 2$;
- **Burst geometry**, where the missing samples occur in bursts of contiguous samples;
- **Random geometry**, where the missing samples do not exhibit any special pattern but are randomly distributed along the original sequence.

A signal with bandwidth b means that the highest normalised frequency in the signal is $b/2$. The nonzero entries of the Γ diagonal define the so-called passband of B [63].

Moreover, note that both sampling and band-limiting are idempotent operations. This means that repeating such operations over the same signal always produce the same result as that obtained from one single operation. Therefore, idempotence allows defining a passband signal x as follows:

$$x = Bx. \tag{4.29}$$

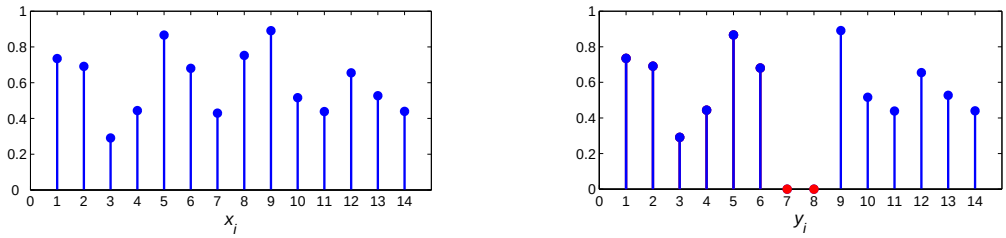
This is useful to formulate the reconstruction problem, specially as it is in subsection 4.2.2.

If the reconstruction algorithm has to solve N equations, *i.e.*, using the whole space of dimension N, then it is called a maximum dimension algorithm. However, if the algorithm only needs to solve n equations (concerning just the unknown samples, for example), then it is called a minimum dimension algorithm. Both cases are based on linear interpolation and operate on a sequence of voice samples of a predefined length, *i.e.*, the number of samples under processing is constant.

Next section describes in detail a maximum dimension reconstruction algorithm that requires low pass band signals of finite energy.

4.2.1 A maximum dimension algorithm

The maximum dimension algorithm under study is the discrete version of Papoulis-Gerchberg algorithm, an iterative linear interpolation algorithm, that is based on [63, 153, 154]. Its aim is to recover missing samples in a finite-length, band-limited data sequence, x, given their positions within the sequence. (Note that in our case, the data sequence of interest is a time segment of a voice signal). Fig. 4.1 shows an example of both original and observed signals, x and y respectively, where y is obtained by setting to zero the 7^{th} and 8^{th} samples of the original signal. Let us call them "erased samples".

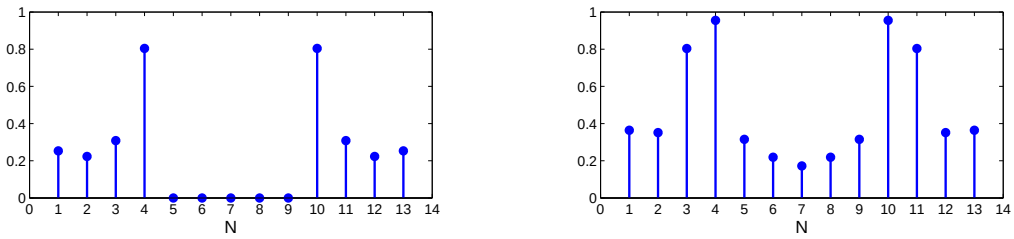


(a) Original signal, x (b) Observed signal, y

Figure 4.1 – Original and observed time-domain signals

In this signal reconstruction algorithm, the unknown data to recover are the values of the erased samples. The known data are the remaining samples present in the observed signal, the position of the erased samples and the bandwidth of the original signal, as given by Eq 4.29. This is equivalent to know the vector y (Fig. 4.1(b)), and the matrices D and B referred to above. Note that, in a practical environment, matrix D is obtained from the received signal by identifying the location of the lost samples and B is obtained from the knowledge of the bandwidth of the original signal, x. For the current case Fig. 4.2 shows the spectral components of both original and observed signals[2].

[2]The DC component is intentionally omitted.



(a) DFT of the original signal, x (b) DFT of the observed signal, y

Figure 4.2 – Spectral components of original and observed signals

Modus operandi

The aim of the reconstruction algorithm is to make the observed signal as close as possible to the original one. By knowing the original signal bandwidth, from which the spectral components are derived, it is possible to compare the observed signal with the original one as the iterative process evolves. Basically, the algorithm filters the observed signal, then extracts the new samples that consequently emerged where they were missing and compose a new observed signal that can again be submitted to such process. This composition is done with the new samples in a sampling-like operation performed in the time domain.

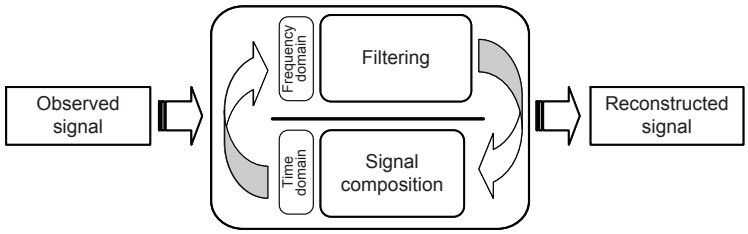


Figure 4.3 – Iterative frequency-domain/time-domain *modus operandi* of the Papoulis-Gerchberg algorithm

This cyclic *modus operandi* is shown in Fig. 4.3. Herein, the main algorithmic steps leading to the reconstructed signal are described in more detail in five steps.

Step 1 – Compute the DFT of y: DFT(y)=Fy

The first step in this algorithm is to transform y into the frequency domain, by computing its DFT, *i.e.*, DFT(y)=Fy. In the subsequent iteration process, the observed signal, y, is subject to several operations. Let us define $y^{(0)}$ as iteration 0 of the reconstructed signal

and $y^{(n)}$ the result of the n^{th} iteration. Iteration 0 is obtained as $y^{(0)} = y = Dx$. As expected, whenever a signal incurs in sharp time-domain variations, such as those originated by loss of samples, this implies changes in the frequency domain. Therefore when losses occur in the original signal, high frequency components appear in the observed signal, y, which lie outside the bandwidth of the original signal. Fig. 4.2(b) shows the result of such operation, where high frequency components (*i.e.*, central components) appeared at locations where originally there were zeros (see Fig. 4.2(a)).

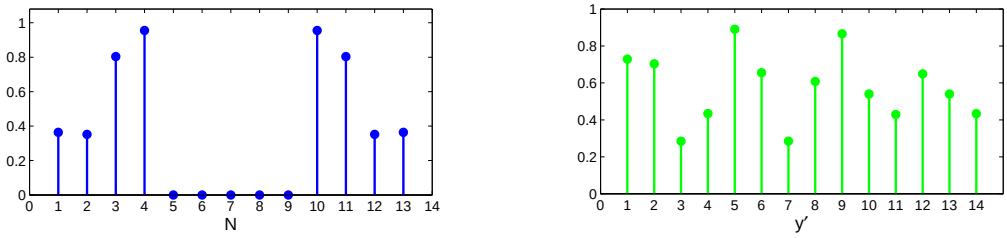
Step 2 – Filter y according to the spectral characteristic of x: $\text{DFT}(y') = \Gamma F y$

The underlying idea behind this process of signal reconstruction is to filter the observed signal y with the same spectral characteristics as those of the original signal, x. Then, the resulting filtered signal, y', is closer to x than y, because its transform domain representation was also approximated to the original one.

Such filtering operation is achieved by left-multiplying matrix Γ by $\text{DFT}(y)=Fy$. It is given by $\text{DFT}(y') = \Gamma F y$. Fig. 4.4 shows both time and frequency domains of the resulting y' signal after filtering the observed signal, y. Fig. 4.4(a) shows the result of this filtering operation, where the undesirable spectral components become zero while the others remain unchanged.

Step 3 – Return $\text{DFT}(y')$ to the time-domain: $y' = F^{-1}\Gamma F y$

The filtered signal y' can now be obtained in the time domain through the the Inverse of the Discrete Fourier Transform (IDFT) of y', $\text{IDFT}(y')$. Note that, as pointed out above, y' is closer to the original signal x than y, *i.e.*, the effect of filtering is to approximate the missing samples towards their original values. Fig. 4.4(b) shows these new samples growing at the sampling instants where their previous values were zero. Note that these are the 7^{th} and 8^{th} samples. However, y' is not the reconstructed signal because this filtering operation also changes the values of the non-corrupted samples of y. All the remaining sample values were also changed, which corrupted the original known samples. Next step restores these ones.



(a) Spectral components of y' after filtering y (b) Time-domain signal, y'

Figure 4.4 – Resulting y' signal after filtering the observed signal, y

Step 4 – Extract the reconstructed samples from the others: $y'' = (I - D)y'$.

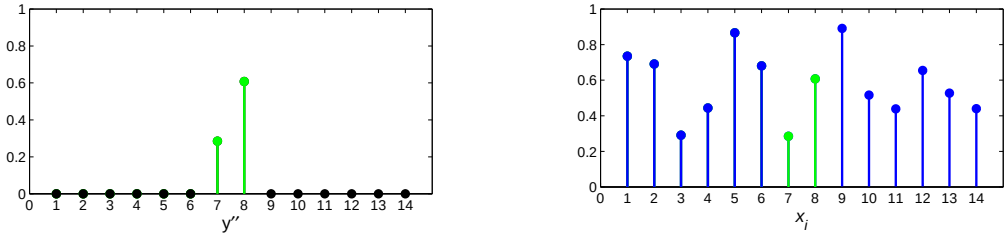
Although the previous process has the advantage of emerging sample values at sampling instants where they were zeros, it also corrupts the good samples of the observed signal y. However, since the time locations of the missing samples are known through matrix D (as in erasure channels, where erasure positions are known), in each iteration is possible to extract only the emerged samples through the following operation $y'' = (I - D)y'$. The output of such extraction process is illustrated in Fig. 4.5(a) where only the emerging samples (relative to the erasure locations) are left and all others are set to zero.

Step 5 – Compose reconstructed signal: $x' = y'' + y$

After the previous step, on the one hand, signal y'' has only non-zero samples at those sampling instants where the missing samples were located in the observed signal. On the other hand, at the remaining sampling instants, the observed signal y contains all non-corrupted samples. This means that signals y and y'' contain non-zero samples at mutually exclusive temporal instants. So the sum of both signals acts as inserting the emerged samples into the observed signal, which results in the first approximation, x', of the original signal, x, and accomplishes the first iteration of the reconstruction process. Fig. 4.5(b) shows the reconstructed signal, x', after the first iteration.

Step 6 – Prepare next iteration: $y = x'$.

If the just reconstructed signal x' has not reached the desired accuracy, it must be considered now as a new observed signal to reconstruct using the same process. For the next iteration, the recently reconstructed signal x' appears as the new observed signal to enter



(a) Signal y'': recovered samples extracted from y'.

(b) Reconstructed signal, x', after the first iteration.

Figure 4.5 – Recovered samples inserted in the observed signal

into a new reconstruction cycle. Therefore, at this step the new observed signal, y, is $y = x'$. Note that, since the non-corrupted samples of the original signal are needed in Step 5 of each iteration, the observed signal used in the first iteration, y, must be stored in memory at the start up of the process, *i.e.*, before Step 1.

After few initial iterations, amplitudes of emerged samples are not yet exactly the same as the original ones, but they tend to the original ones as the iterative process converges to a more accurate solution. Fig. 4.6 shows the described steps.

The algorithmic steps described above can be defined by a sequence of algebraic expressions, which are the basis for software implementation of the reconstruction method. Taking the reverse path from Step 5 to Step 1, the following expressions fully describe the reconstruction algorithm.

$$\begin{aligned} x' &= y'' + y \\ x' &= (I - D)y' + y \\ x' &= (I - D)F^{-1}\Gamma F y + y \end{aligned} \tag{4.30}$$

From (4.30) it is possible to obtain the following expression for the reconstructed signal in the $(k+1)^{th}$ iteration

$$x'^{(k+1)} = (I - D)Bx'^{(k)} + y, \tag{4.31}$$

where $x'^{(k)}$ represents the reconstructed signal at iteration k.

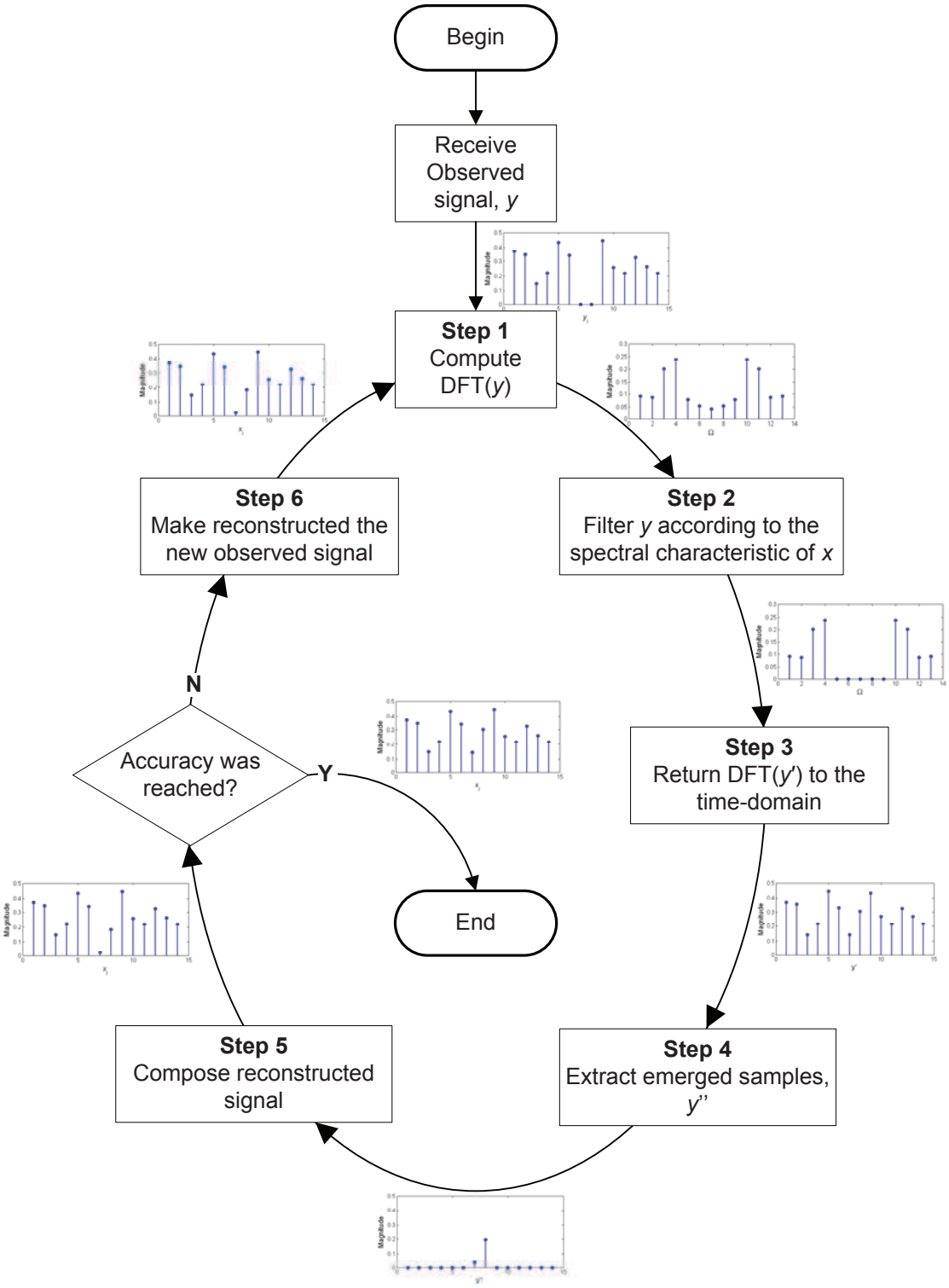


Figure 4.6 – Cyclical iterative *modus operandi* of the Papoulis-Gerchberg algorithm

It is possible to isolate the iteration matrix by visual inspection. It is of the form

$$A = (I - D)B. \tag{4.32}$$

In order to improve convergence, a relaxation constant, μ is used with similar role as w in Eq. 4.27. Thus Eq. 4.31 becomes,

$$x'^{(k+1)} = (I - \mu D)Bx'^{(k)} + \mu y \tag{4.33}$$

and the iteration matrix is given by

$$A_\mu = (I - \mu D)B. \tag{4.34}$$

For an effective use of such a constant, the values belonging to the interval $]0, 2[$ must be used. More precisely, the optimum value of μ is given by

$$\mu_{opt} = \frac{2}{2 - \lambda_{max}} \tag{4.35}$$

where $\lambda_{max} = \rho(S_1)$ with S_1 being $S_1 = B(I - D)B$ [63].

The described algorithm converges if the density of sampling, s/N is greater than the signal bandwidth, *i.e.*, $s/N > q/N$. Thus, convergence can be guaranteed by reducing the number of missing samples and/or the bandwidth.

An important issue concerning convergence is the error pattern geometry (*i.e.*, the location of missing samples) which influence the asymptotic convergence rate of the algorithm, for a given density. The convergence rate partially depends on matrix B and on the error geometry. In fact, for low pass signals, the best possible error patterns are those in which the missing sample time positions are equidistant. The worst possible geometry is that of contiguous missing samples. In the middle there is the random geometry [155]. This issue is important in order to obtain a well-conditioned problem. Due to the fact that this is a maximum dimension method, the possibility of low convergence rates and the computing resources required per iteration (essentially a pair of FFTs) are disadvantages of this approach. Since this is a maximum dimension problem, it is expected to exhibit a relatively low convergence rate due to the enormous computing resources required. This is further discussed in Section 4.3. Next section describes a minimum dimension algorithm.

4.2.2 A minimum dimension algorithm

As mentioned in previous sections, a minimum dimension algorithm is characterised by a system of only n equations corresponding to the n unknown samples. This subsection describes a minimum dimension algorithm, which also requires band-limited signals of finite-dimension similarly to the Papoulis-Gerchberg algorithm described in subsection 4.2.1, *i.e.*, $x = Bx$ (Eq. 4.29) must be valid.

To establish the basic concepts of this algorithm, the specific case of an original signal x_i with length $N = 5$ is used, *i.e.*, $x_i = x_1, x_2, x_3, x_4, x_5$. For this signal, Eq. 4.29 becomes

$$\begin{aligned} x_1 &= b_{11}x_1 + b_{12}x_2 + b_{13}x_3 + b_{14}x_4 + b_{15}x_1 \\ x_2 &= b_{21}x_1 + b_{22}x_2 + b_{23}x_3 + b_{24}x_4 + b_{15}x_2 \\ x_3 &= b_{31}x_1 + b_{32}x_2 + b_{33}x_3 + b_{34}x_4 + b_{15}x_3 \\ x_4 &= b_{41}x_1 + b_{42}x_2 + b_{43}x_3 + b_{44}x_4 + b_{15}x_4 \\ x_5 &= b_{51}x_1 + b_{52}x_2 + b_{53}x_3 + b_{54}x_4 + b_{15}x_5 \end{aligned} \quad , \tag{4.36}$$

where b_{ij} are the elements of the matrix B.

For reconstruction purposes let us assume that the 2^{nd} and 4^{th} samples of x_i are lost. Then the set of equations 4.36 are limited to those including the lost samples. In each of these equations, we are interested in separating the right side terms containing unknown samples (x_2, x_4) from those containing the known ones. This yields

$$\begin{aligned} x_2 &= b_{21}x_1 + b_{22}x_2 + b_{23}x_3 + b_{24}x_4 + b_{25}x_5 \\ x_4 &= b_{41}x_1 + b_{42}x_2 + b_{43}x_3 + b_{44}x_4 + b_{45}x_5 \end{aligned} \tag{4.37}$$

which is equivalent to

$$\begin{bmatrix} x_2 & x_4 \end{bmatrix} = \begin{bmatrix} b_{22} & b_{24} \\ b_{42} & b_{44} \end{bmatrix} \begin{bmatrix} x_2 \\ x_4 \end{bmatrix} + \begin{bmatrix} b_{21} & b_{23} & b_{25} \\ b_{41} & b_{43} & b_{45} \end{bmatrix} \begin{bmatrix} x_1 \\ x_3 \\ x_5 \end{bmatrix} \tag{4.38}$$

Let us denote by u_i the subset of the original signal x_i which contains the unknown values. In this case, $u_i = \{x_2, x_4\}$ is of cardinality n=2. Also, let us define $U = \{i_1, \cdots, i_n\}$ as the set of subscripts of n unknown samples in x_i. In the present case, $U = \{2, 4\}$. Therefore,

equations 4.38 can be written as

$$u_i = \sum_{j \in U} b_{ij} x_j + \sum_{j \notin U} b_{ij} x_j; \quad \text{for all } i \in U \tag{4.39}$$

or, in matricial form

$$u = Su + h, \tag{4.40}$$

where S is a $n \times n$ principal sub-matrix of B, as defined in 4.38, and h is the $(N-n)$-dimensional vector in the second sum of Eq. 4.39, which is a linear combination of the known samples of x_i. The conditions under which these equations provide a solution for u can be found in [149]. Eq. 4.40 can be solved by using two techniques: matricial direct computation and iterative computation. With the former case in view, Eq. 4.40, becomes successively equivalent to

$$\begin{gathered} u = Su + h, \\ u - Su = h, \\ Iu - Su = h, \\ (I - S)u = h, \\ (I - S)^{-1}(I - S)u = (I - S)^{-1}h, \\ u = (I - S)^{-1}h. \end{gathered} \tag{4.41}$$

This result is valid, providing that $(I - S)^{-1}$ exists. Thus Eq. 4.40 has a unique solution regardless the number and distribution of the lost samples.

If Eq. 4.40 is solved through an iterative process, then it suggests the following form where a non-relaxation method is used.

$$u^{(k+1)} = Su^{(k)} + h. \tag{4.42}$$

Then $u^{(k)}$ is obtained at iteration k and the solution is given by the limit

$$u = \lim_{i \to \infty} u^{(i)}, \tag{4.43}$$

regardless of $u^{(0)}$. The condition $\rho(S) < 1$ guarantees that such limit exists, where S is the system matrix [149].

As it was pointed out before, two different techniques can be used to solve Eq. 4.40: direct calculation and iterative methods. Direct calculation of u, as given in Eq. 4.41, has the

advantage of being done in one single step, providing that $(I - S)^{-1}$ exists. In practice, there are several factors which may lead to serious difficulties in calculating the inverse of $I - S$. For example, if one of the eigenvalues of S is close enough to the unity, then computation of $(I - S)^{-1}$ may become very difficult, or even impossible, leading to an ill-conditioned problem. In such cases, an iterative method may be used to circumvent this difficulty and to find an accurate approximation for solution, u. Despite the fact of having an ill-conditioned problem, in the case of direct calculation such problem is impossible to solve, whereas in the case of iterative methods an approximation is always possible to be found, though its accuracy may not be very high.

The eigenvalues of the system matrix S depend on the distribution of the missing samples. In particular, its spectral radius is more likely be unitary for burst distributions rather than for equidistant missing samples [150]. In the case of signal reconstruction it is interesting to note that, if the distribution of the missing samples u_i=$\{u_{i_1}, u_{i_2}, \cdots, u_{i_p}\}$ is equidistant by some fixed integer $m \geq 1$, that is, $u_i = \{im\}$, $i = 1, \cdots, p$, then the eigenvalues of S, $\lambda_i(S)$, have a lower bound given by $(\lfloor rm \rfloor + 1)/m$ and an upper bound given by $\lceil rm \rceil / m$, *i.e.*,

$$\frac{\lfloor rm \rfloor}{m} \leq \lambda_i(S) \leq \frac{\lceil rm \rceil}{m} \leq 1, \tag{4.44}$$

where $\lfloor rm \rfloor$ denotes the greatest integer less than or equal to rm and $\lceil rm \rceil$ denotes the smallest integer equal or greater than rm. In the particular case of $r = \lfloor rm \rfloor / m$, the eigenvalues of S are all the same, $\lambda_i(S) = r$, $\forall_i$. In such case $S = rI$. In the particular case in which the missing samples are equidistant, S becomes Toeplitz.

Given the above analysis, it figures out that if it is possible to select the gap between missing samples, it will be possible to control the problem and put it into a well-conditioning point. In a real system, where voice is transmitted in packets, the value of this gap can be induced by interleaving the originally contiguous samples. In this way, contiguous samples are spread out along several different packets which means that the loss of one packet represents the loss of originally equidistant samples, as shown in the Fig. 4.7. The first raw represents the original sequence of samples. The second raw shows how the samples are packetised by interleaving them: first sample is put in the first position of the first packet, the second sample is put into the first position of the second packet, ..., and the

fourth sample is put into the first position of the fourth packet. Then this round-robin filling process restarts by putting the remaining samples into the remaining positions of the four packets. The third raw represents the loss of the third packet (in grey) and the fourth raw shows that, if samples are reordered to be like in the original sequence, the missing samples become equidistant. In this case an interleaving factor, m, of $m = 4$ is used since the samples which position are multiple of 4 are picked up to be put in the same packet.

As it can be seen, loosing a packet (grey samples) after samples have been interleaved means to lose equidistant samples after a reordering operation was performed in the receiver. Then it is possible to put $\lambda_i(S)$ close to either r or its multiples, regardless of the number of missing samples. By using an appropriate choice of the oversampling and interleaving factors r and m (*i.e.*, such that $m \times r$ is an integer) respectively, it is possible to put $\lambda_i(S)$ less enough than unity in order to control the reconstruction accuracy and processing speed.

In VoIP context, the use of an adequate interleaving factor m at the source, not only

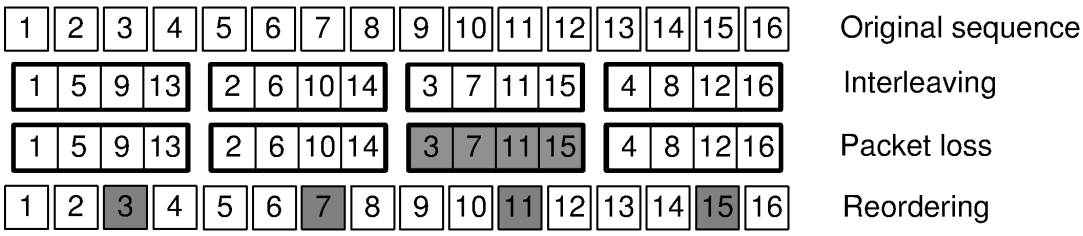


Figure 4.7 – Interleaving as a mean to make lost samples equidistant

makes such a signal more robust against possible degradations by transforming burst errors in equidistant ones, but also makes reconstruction easier because it leads to a well-conditioned problem. Therefore, when a packet is lost with n voice samples in its payload, this leads to a reconstruction problem where missing samples are equidistantly distributed, separated by m-1, and the matrix S of the resulting reconstruction problem is of dimension $n \times n$ ($S_{n \times n}$).

Fig. 4.8 shows the maximum and the minimum eigenvalues of S as a function of the interleaving factor, m, for a given bandwidth (defined by r), as given by Eq. 4.44. In this

case, $r = 0.6$. As the figure shows, greater values of m lead to better conditioned problems because λ_{max} decreases as m increases. Also, when the product $r \times m$ is an integer, all eigenvalues are equal since they are $\lambda_i(S) = r$, as stated before. In Fig. 4.8, this occurs for $m = 5$ and $m = 10$.

Considering real time implementation issues, while the solution of Eq. 4.40 must be found in real time to be effective, the possibility to have a pool of different dimension system matrices S, previously calculated and stored in memory, turns the whole reconstruction process more expedite. Furthermore, its calculation may become as trivial as $S = rI$, as stated before.

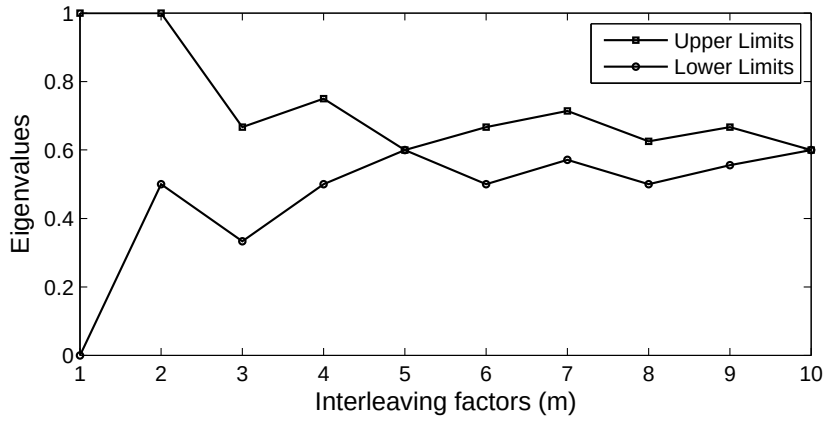


Figure 4.8 – Spectral radius *vs.* interleaving factor (r=0.6).

4.3 Simulation results

In this section, the reconstruction algorithms previously described are evaluated through simulation and the results are analysed and discussed. The simulation study is also aimed to provide a deeper understanding of the most important factors influencing the problem of signal reconstruction and to show how greater the performance of the minimum dimension algorithm is when compared with the maximum dimension one.

In the case of the Papoulis-Gerchberg algorithm, it is important to analyse the factors that influence the conditioning of the reconstruction problem and how they influence its

solution. These factors include the spectral radius of the iteration matrix, the distribution of the missing samples (error geometry) and the signal bandwidth. The performance is evaluated by measuring the number of iterations necessary to reach the solution, the percentage of lost samples, the spectral radius of the iteration matrix and the RMSE between the original and reconstructed signals. The RMSE is given by the expression 4.45, where $x[i]$ is the original signal, $x'[i]$ the reconstructed signal and N the sequence length.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x[i] - x'[i])^2}. \tag{4.45}$$

The stop criterion was defined as an upper bound for residual error between consecutive iterations, given by

$$residual = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x'^{(k+1)}[i] - x'^{(k)}[i])^2} \leq 10^{-8}, \tag{4.46}$$

where $x'^{(k)}[i]$ is the i^{th} sample of the reconstructed signal at iteration k. In this case, imposing a certain maximum residual value means to stop the reconstruction process when the achieved gain (*i.e.* accuracy) does not worth the additional computational effort. In the case of random geometry tests, *i.e.*, those where the distance between voice samples is random, each new simulation run uses more missing samples than the previous one, which in turn acts as a seed in order to guarantee an increasing spectral radius over successive runs. The same original voice signal was used in all the tests. In all experiments, a voice signal with $N = 256$ samples was used. In the case of the Papoulis-Gerchberg algorithm, three different error distributions, referred to as interleaved, random and burst geometries were used. These experiments run on a computer equipped with an Intel T2300@1.66 MHz processor and 1.5 GB RAM. Also, two different signal bandwidths were used, defined by two different factors, r: $r = 0.8$ and $r = 0.6$.

4.3.1 The maximum-dimension algorithm: discussion

The performance of Papoulis-Gerchberg algorithm was evaluated by carrying out three types of tests intended to find out how the spectral radius of the iteration matrix, the error geometry of lost samples and the signal bandwidth influence the convergence of the algorithm. Next three subsections discuss these tests and results.

The spectral radius of the iteration matrix

This test is intended to evaluate the influence of the spectral radius of the iteration matrix of Eq. 4.32, $\rho(A)$, in the algorithm convergence. In the experiments, the oversampling factor was set to $r = 0.6$ and the relaxation constant set to $\mu = 1$. The percentage of missing samples varied from 0.4% to 25%. Fig. 4.9 shows the number of iterations necessary to obtain a residual error less than 10^{-8}, as a function of the spectral radius, $\rho(A)$. As one can observe, as the spectral radius of the iteration matrix increases, the number of iterations also increases, suggesting an exponential function of the spectral radius. In fact, it increases in inverse ratio to the logarithm of the spectral radius, as explained below.

The absolute error of the approximated solution (not the residual error) obtained after iteration $k+1$ is bounded by

$$\|e_{k+1}\| \leq \lambda_{max}^{k} \|e_1\| \tag{4.47}$$

where λ_{max} represents the maximum eigenvalue of the system matrix and e_1 represents

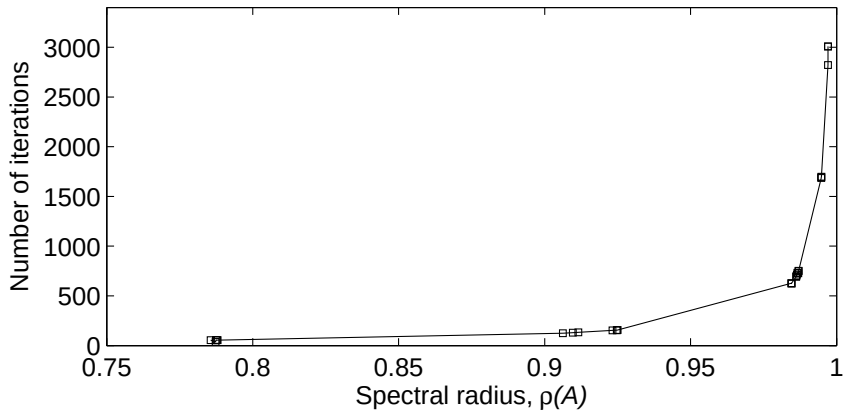


Figure 4.9 – Number of iterations to obtain residual error $<10^{-8}$ *vs.* spectral radius, $\rho(A)$ (*r=0.6)*

the error in the first iteration [149].

On the one hand, this expression shows that in order to attain a given error, higher values of λ_{max} imply higher number of iterations, since λ_{max} is less than 1, k is greater than 1 and $\|e_1\|$ is constant. It is also possible to see that, if $\lambda_{max} = 1$, the error after iteration $k+1$ will never decrease below to that of the first iteration, thus the algorithm does not

converge.

On the other hand, it is possible to see why the number of iterations increases in inverse ratio to the logarithm of the spectral radius. By applying the logarithmic function to both members of Eq. 4.47 it is possible to obtain $k \leq \log(C)/\log(\lambda_{max})$, where C is a constant given by $C = \|e_{k+1}\|/\|e_1\|$. As λ_{max} reaches 1, $\log(\lambda_{max})$ reaches 0 which makes k tend rapidly to infinity since $\log(C)$ is a constant. It is also possible to derive the number of iterations, k, by fixing the upper bound of the error in the $(k+1)^{\text{th}}$ iteration, given that the error in the first iteration is known.

In Fig. 4.9 it is evident that $\rho(A) = 1$ leads to a non-convergence situation. Therefore, an important conclusion is that lower spectral radii lead to better convergence and spectral radii near to 1 can easily turn the problem into ill-conditioned making convergence difficult or even impossible. In this case an acceptable solution cannot be found.

Error geometry

This test is intended to evaluate the influence of the missing samples distribution on convergence and to identify the break even points, *i.e.*, the maximum percentage of missing samples for which the algorithm still converges. Note that each break even point also corresponds to a specific spectral radius because this is implicitly defined by the missing samples. In the experiments, the values $r = 0.8$ and $\mu = 1$ were used. Interleaved, random and burst distributions with loss percentages ranging from 1% to 50% were used. To determine the break even points, spectral radii close to 1 were used. Fig. 4.10 shows the spectral radius as a function of the percentage of missing samples, for the three error geometries under study and $r = 0.8$. As the figure shows, the spectral radius depends on two factors: the percentage of missing samples –which can be seen taking one of the three lines–, and the error geometry –which can be seen by comparing the three lines–. In regard to the percentage of missing samples, one can observe that the spectral radius increases as more samples are missing in the signal. This is common to all the three geometries, which exhibit the same behaviour. In the case of different error geometries, one can also see that for a given spectral radius, the interleaved geometry is the one which tolerates more missing samples and the burst geometry is the one that tolerates less missing

samples, since just two or three missing samples are enough to make the spectral radius be close to 1. The behaviour of the random geometry is between the other two. From another point of view, for the same number of missing samples, the interleaved geometry exhibits a lower spectral radius than the others and the burst geometry exhibits the greatest spectral radius. These results lead to the conclusion that interleaved geometry is the error pattern that tolerates a greater percentage of missing samples in the signal.

Fig. 4.11 shows the break even points, for each type of error geometry defined by either

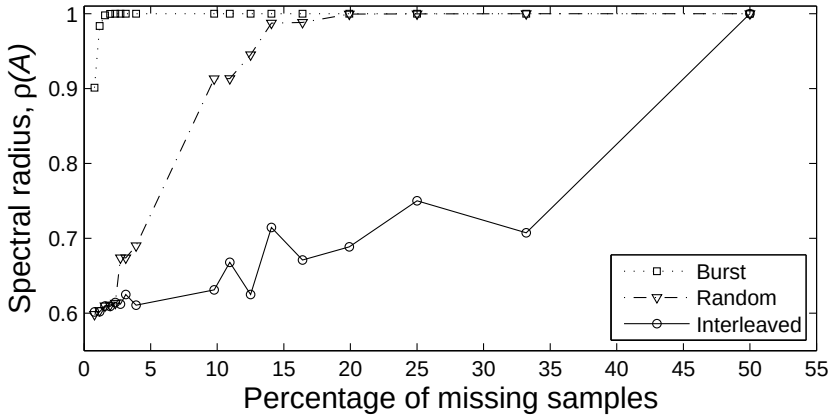


Figure 4.10 – Spectral radius *vs.* percentage of missing samples, for three error geometries (r=0.8)

the percentage of missing samples and its corresponding spectral radius. Note that these are the maximum values for which the algorithm still converges to a unique solution.

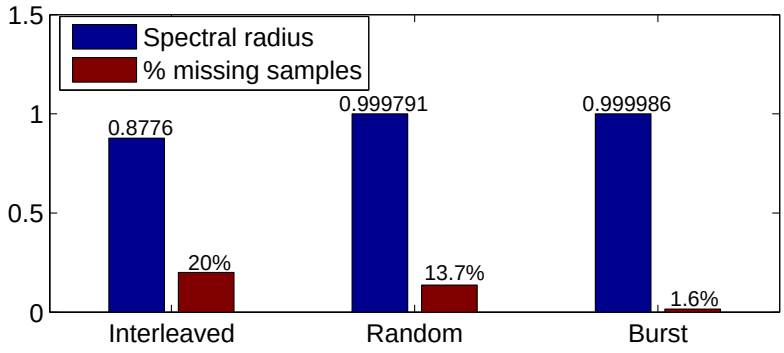


Figure 4.11 – Break even points for each geometry (r=0.8)

These results show that break even point positions vary according to the error geometry.

In the case of the interleaved geometry, the maximum spectral radius that still leads to a well-conditioned problem is 0.8776. It corresponds to an interleaving factor of missing samples of $m = 5$ (*i.e.*, 1 out of 5) thus to 20% of missing samples. Other spectral radii between 0.8776 and 1 are possible to achieve, but the next value for the interleaving factor is four ($m = 4$), which leads to a spectral radius of 1, thus to a ill-conditioned problem. In the case of the random geometry, the maximum spectral radius that still leads to a well-conditioned problem is 0.999791 corresponding to lose 13.7% of the signal samples. Finally, the worst situation is the burst geometry in which the maximum possible spectral radius is 0.999986 corresponding to lose 1.6% of the signal samples. Overall, these results confirm that the interleaved error geometry is the most tolerant to losses in the sense that more signal samples may be lost before the problem becomes ill-conditioned. On the opposite side, the burst error geometry was found to be the less tolerant to losses.

Influence of the signal bandwidth

This test is aimed to find out the influence of the signal bandwidth on the algorithm convergence. The test is similar to that concerning the spectral radius of the iteration matrix (depicted in Fig. 4.9) except the signal bandwidth which was decreased through the oversampling factor, now set to $r = 0.4$. The results in Fig. 4.12 show that for spectral radii less than 0.8 the number of iterations required to converge is significantly reduced as compared with higher spectral radii. Therefore, as expected, faster convergence is achieved for lower signal bandwidth.

It has been seen that bandwidth influences the convergence. Another relevant issue is to find out how the break even points are affected by decreasing the signal bandwidth. For this, the oversampling factor was reduced from $r = 0.8$ (test to error geometries, Fig. 4.11) to $r = 0.4$. Fig. 4.13 shows that break even points are achieved at higher values than in the case of $r = 0.8$ (Fig. 4.11). This means that a greater percentage of missing samples is allowed in signals with lower bandwidth without reaching the non-convergence boundary. For the interleaved geometry, the maximum spectral radius that still guarantees convergence is 0.5. Since the respective interleaving factor is $m = 2$, then 50% of samples are allowed to be lost in this case. Comparing with results obtained in the tests concerning the error geometry (Fig. 4.10) and the spectral radius of the iteration

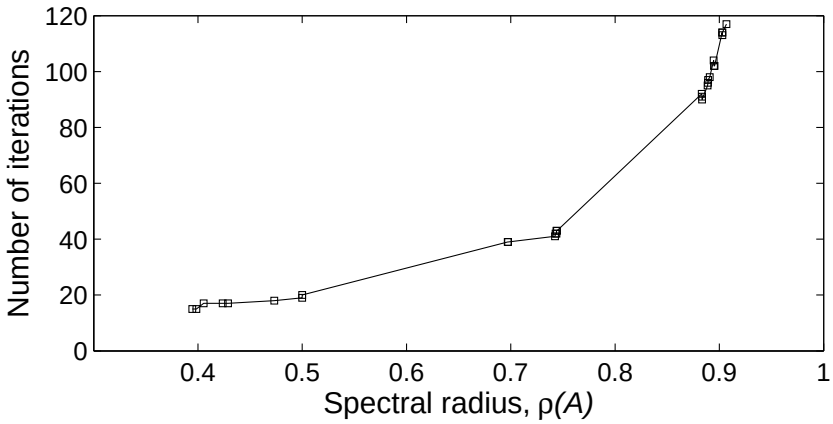


Figure 4.12 – Number of iterations to obtain residual error $<10^{-8}$ *vs.* spectral radius, $\rho(A)$ (*r=0.4)*

matrix (Fig. 4.9), where the signal bandwidth was greater ($r = 0.8$), this corresponds to a significant improvement in tolerance to loss of samples. Note that in the previous case the maximum sample loss rate was just 20%. The same behaviour occurs for the random and burst error geometries. In the case of random losses, for the maximum spectral radius that still leads to a convergent situation (*i.e.*, 0.999991), 50% of missing samples are still allowed against 13.7% in the case of $r = 0.8$. In the case of error bursts, for the maximum allowed spectral radius of 0.999968, it is possible to have 3.9% of missing samples against 1.6% in the case of $r = 0.8$.

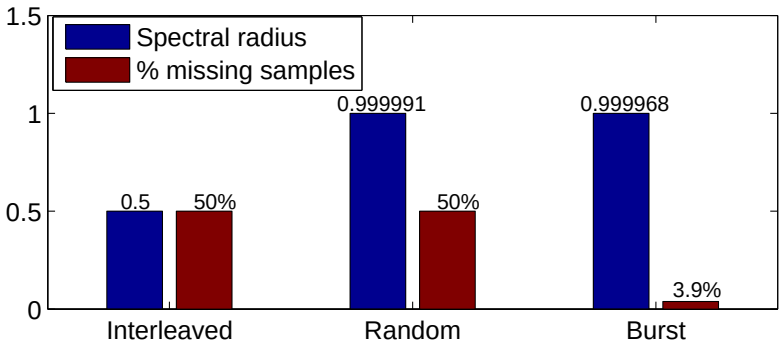


Figure 4.13 – Break even points for each geometry (*r=0.4)*

These results show that the signal bandwidth influences the convergence rate. In our case, where signals are low pass signals, we conclude that a lower signal bandwidth leads to

greater convergence rates. Also, the interleaved geometry is shown to be more tolerant to losses, which leads to the conclusion that a mechanism that interpolates voice samples in the source is more adequate to improve error robustness at the source and to ease signal reconstruction at the destination when the method is used, exclusively. In chapter 6 (section 6.4) a technique that uses the Papoulis-Gerchberg algorithm combined with the method therein described permits to give more robustness to the reconstruction when small bursts of erased samples occur.

4.3.2 The minimum dimension algorithm: discussion

The experiments described in this subsection are intended to evaluate and compare the performance of the Papoulis-Gerchberg (PG) method with that of the minimum dimension method using both the iterative (MD Iterat) and direct computation (MD Direct) variants. The performance metrics used in the study were the processing time obtained from Matlab© and the RMSE between the original and the reconstructed signals. Since the spectral radius plays an important role in the reconstruction accuracy and processing time, the dependence on the number of unknown samples was also studied. After analysing the results obtained in section 4.3.1, where low pass signals were used, interleaved error geometry appears as the most realistic in the case of a real working system. This is the reason by which experiments of this subsection concern only interleaved geometry.

Spectral radius

Fig. 4.14 shows the dependency of the spectral radius from the percentage of missing samples for the various reconstruction methods. It is evident from the figure that the spectral radius increases with the number of missing samples, which means that in all methods more missing samples tend to result in ill-conditioned reconstruction problems, when these interpolation algorithms are used. This is in line with the results of subsection 4.3.1. In fact, the three curves concerning $r = 0.8$ (PG, MD Direct and MD Iterat) coincide and the three curves representing the same methods but for $r = 0.6$, also coincide.

Another important conclusion is that the spectral radius of the system matrix is independent from the reconstruction method for both oversampling factors $r = 0.8$ and $r = 0.6$. Moreover, it can be seen that greater bandwidth (*i.e.*, greater r) implies greater spectral radii, which makes one to expect more processing time in the respective reconstruction. This is also in line with the conclusions of subsection 4.3.1. Note that coincident lines in the figure means that for each value of r, the spectral radii are the same for all methods.

Accuracy

Fig. 4.15 shows how the RMSE between reconstructed signal and the original one depends on the number of missing samples. The break even points are also represented in the figure, separating the well-conditioning region (left side) from that of ill-conditioning (right side). One can also observe that both iterative methods (PG and MD Iterat) achieve the same RMSE with the critical point occurring when the spectral radii, $\rho(A)$ and $\rho(S)$, of the system matrices, A and S, are close to 1 ($\rho(A) = \rho(S) \cong 1$; see figure 4.15). $\rho(A)$ denotes the spectral radius of the maximum dimension algorithm matrix and $\rho(S)$ denotes the spectral radius of the minimum dimension algorithm matrix. For both methods, these spectral radii have the same value and $\rho(A) = \rho(S) = 0.88$ corresponding to 20% of missing samples with an interleaving factor $m = 5$.

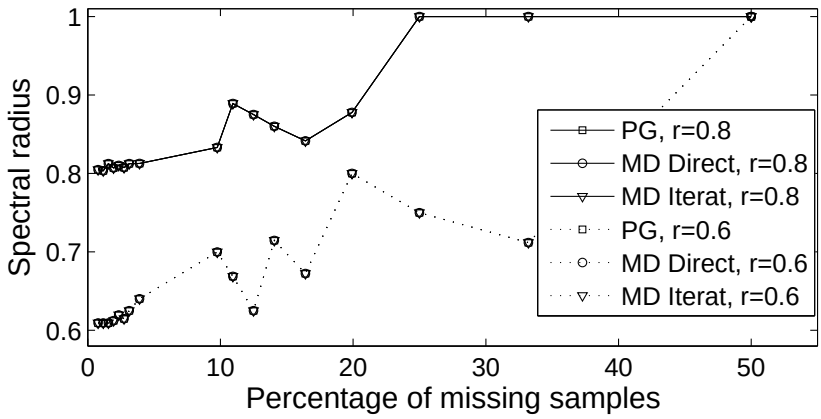


Figure 4.14 – Spectral radius *vs.* missing samples for each method and oversampling factor, r

Important to notice is that for small percentages of missing samples (below 25%), the direct computation variant of the minimum dimension problem (MD Direct) provides more accurate reconstructed signals than either maximum or minimum dimension iterative methods. (The same accuracy is obtained from both iterative methods when the number of missing samples is low). For larger number of missing samples ($\gtrsim 25\%$, in this case), iterative methods exhibit slightly higher reconstruction accuracy. Therefore, when the problem is well-conditioned, direct variant computation is more suitable, whereas in the case of a ill-conditioned problem, iterative methods are preferable.

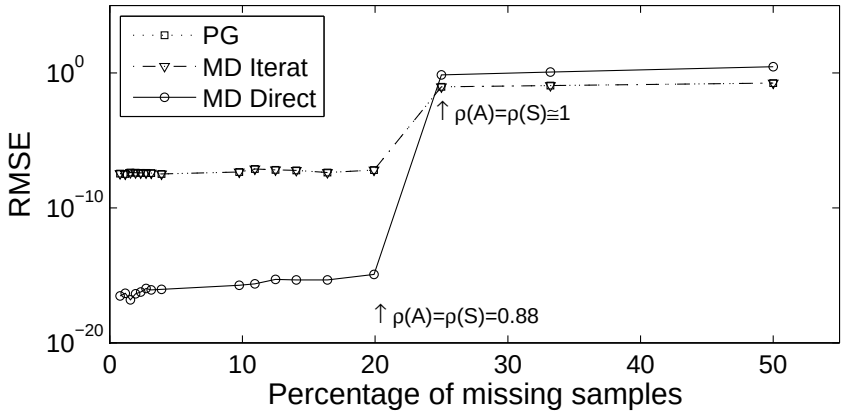


Figure 4.15 – RMSE between reconstructed and original signals *vs.* percentage of missing samples for maximum and minimum dimension algorithms and r=0.8

Fig. 4.16 shows similar results as in Fig. 4.15, except that the signal bandwidth, r, is lower, in this case. The results in this figure confirm that, in the case where the number of missing samples is small, the direct variant of the minimum dimension algorithm (MD Direct) gives better reconstruction accuracy than iterative variants for both algorithms. However, for large number of missing samples, iterative variants exhibit slightly better reconstruction accuracy. The break even points are the same for both algorithms but they are now shifted to the right, which means that more missing samples are allowed. In this case, it corresponds to a spectral radius $\rho(A) = \rho(S) = 0.71$ and 33.3% of missing samples ($m = 3$). This result corroborates what was saw in subsection 4.3.1.

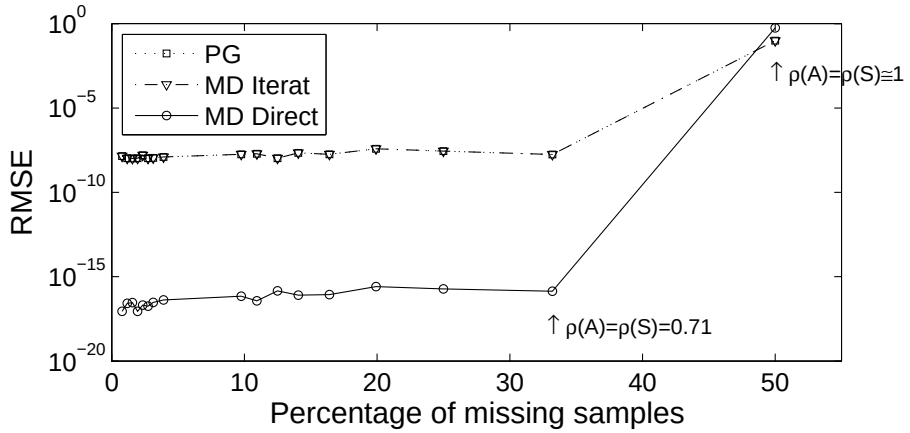


Figure 4.16 – RMSE between reconstructed and original signals *vs.* percentage of missing samples for a maximum and minimum dimension algorithms and r=0.6

Computation time

The computation time spent by the reconstruction algorithms are shown in Fig. 4.17 and Fig. 4.18, for the case of r = 0.8 and r = 0.6, respectively. Both maximum and minimum dimension algorithms and the iterative and direct computation variants of the minimum dimension were evaluated. As it can be seen in both figures, for a small number of missing samples, direct computation of the minimum dimension problem is the fastest one and a lower bandwidth signal leads to smaller computation time, particularly when using an iterative method. However, for a large number of lost samples the direct method is more time consuming. This result is in line with that of RMSE, as shown in Fig. 4.15: for low percentages of missing samples, direct computation is better whereas for high percentages of missing samples, iterative processes are better.

The processing time of the Papoulis-Gerchberg algorithm is always higher than that of the minimum dimension one, regardless of its variant, either iterative or direct computation. This result reveals the importance of using a minimum dimension algorithm. In fact, as it was been said in subsection 4.2.2, the reduced number of equations when compared with the maximum dimension algorithm make to be expected the reduction of the computation time. However the difference between them decreases when the number of missing samples increases. This is because in such case the problem dimension in the minimum dimension method approximates the dimension of the Papoulis-Gerchberg.

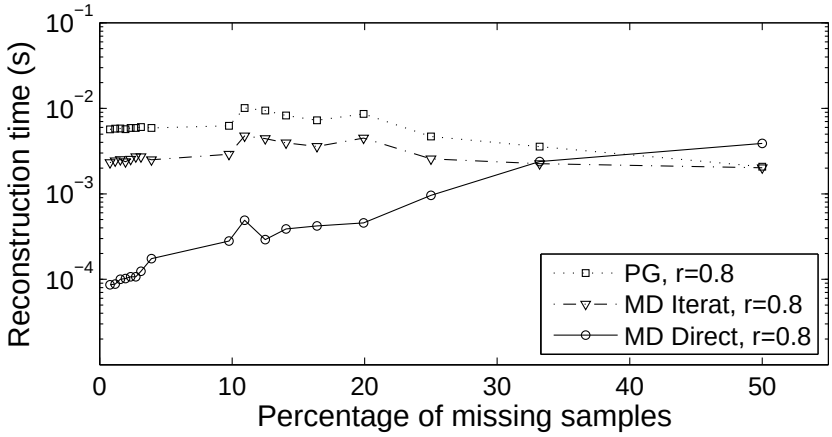


Figure 4.17 – Computation time of reconstruction; r=0.8

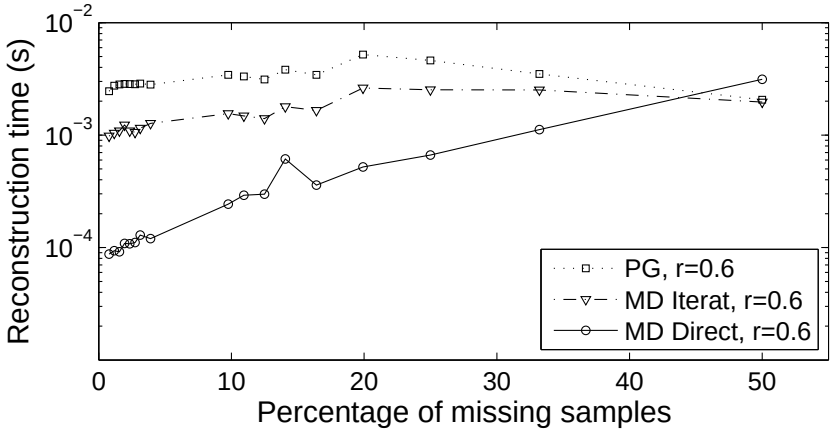


Figure 4.18 – Computation time of reconstruction; r=0.6

As in the case of the maximum dimension algorithm, an interleaved geometry to apply to a voice signal at the source is shown to be more tolerant to losses, which leads to the conclusion that such a mechanism is more adequate to improve error robustness and to ease signal reconstruction.

4.4 Conclusion

This chapter presented a study about useful solutions for signal reconstruction, to overcome problems due to lost samples in voice signals. Special emphasis was given to a detailed description and comparison of two linear interpolation algorithms for voice signal

reconstruction.

In a first stage, maximum and minimum dimension as well as iterative and direct computation concepts and algorithms were presented, from which three reconstruction methods were identified. Factors that influence the accuracy of the methods were identified and confirmed by a simulation study. These factors are the spectral radius of the system matrix, the error geometry of the lost samples and the signal bandwidth. The achieved results permitted to conclude that these algorithms are better suitable to reconstruct lost samples when these present an interleaved geometry. Interleaving voice samples at the time when they are packetised was foreseen as a robust method to ease reconstruction of the signal.

In a second stage, attention was focused on the interleaved geometry to compare the performance obtained by the three identified methods. It was possible to conclude that bellow a certain percentage of losses, a direct method gives better accuracy and lower computation times. Above a certain percentage of losses, iterative methods are more suitable, since they give better accuracy and lower computation times.

The possibility to put the problem in a well-conditioning point *a priori* by choosing an adequate combination of the oversampling and interleaving factors, r and m, respectively, as well as having a pool of pre-calculated matrices, S, permits to consider these methods suitable to implement in a VoIP system.

5 A Practical Model for Voice Quality Evaluation

This chapter describes a research study that was carried out in order to devise a practical voice quality evaluation method. It relies on both the ITU-T Rec. G.107 and the ITU-T Rec. G.108 and uses the PESQ as the reference calibration method.

This research is the second stage of the work that begun with the study of the methods presented in chapter 3. This work is also part of a R&D project addressing the evaluation of voice quality developed in collaboration with PTIn. The E-Model is the basis of the research work described in the following sections.

5.1 Methodology

To begin this study and efficiently carry out the aimed experiences, a methodology is established, encompassing the following phases:

- The choice of standard methods to base the model to derive, as well as the standard method to validate such model;
- The use of a call quality monitoring system to provide input parameters as well as evaluation results to test the derived model;
- The accomplishment of preliminary experiences using the base model to get practical about its behaviour under different conditions.

5.1.1 E-Model: the base model

A comparative evaluation study of the most recent voice quality evaluation methods was done in order to base our prototype. The first methods to consider as candidates are those that give the true opinions of users: the subjective methods as described in the ITU-T Rec. P.800. However, implementing such methods is not realistic for our case since they are not compatible with the needs of the repeatability demanded by a real time production system among other logistic problems already mentioned in section 3.3.

The PESQ method emerges as a good alternative since it does not require to recruit users to carry out the evaluation experiences and is, currently, one of the most widely accepted methods. However, since this is an intrusive method, it is difficult to use for evaluation of a system in production.

The ITU-T Rec. P.563 method must also be considered since it is a non-intrusive method, which is easy to be implemented in a real-time system. However, the requirement to avoid dealing with intellectual property rights led to seek for alternative solutions.

The E-Model, proposed in the ITU-T Rec. G.107, and supported by the practical implementation guide described in the ITU-T Rec. G.108, is the model that presents the best option due to three main reasons: i) there are no claimed intellectual property rights, ii) it is a non-intrusive method and iii) it does not need to recruit people, which is a great advantage, given the requirements of the current project. An additional advantage, not strictly necessary but an important asset, is that this method gives a conversational MOS (MOS_{CQS}), since it takes into consideration delay factors. The drawback of this method is the need to know the characteristic parameters of each equipment that constitute the end-to-end circuit and the circuit as a whole. For the current project, we were aware, *a priori*, that the knowledge of the characteristics of all the equipments would not be possible. However, since the model permits to use default values (as presented in Table 3.13), in case the real ones are unknown, it is possible to circumvent this problem at the expense of some acceptable accuracy decrease.

Table 5.1 – Summary of advantages and disadvantages of the candidate methods

Candidate methods	Advantages	Disadvantages
MOS (P.800)	True quality evaluation	People recruitment Expensive Time-consuming Non-repeatable
PESQ (P.862)	No people recruitment Widely accepted	Intellectual property rights Intrusive
P.563	No people recruitment Non-intrusive	Intellectual property rights
E-Model	No people recruitment Non-intrusive No intellectual property rights Conversational MOS (MOS_{CQS})	Need to know characteristics of each equipment and circuit

Table 5.1 summarises the discussed advantages and disadvantages of the referred methods. Given the E-Model advantages and the possibility to circumvent the disadvantages, it was chosen as the basis to derive the proposed evaluation model. A reference method has been chosen for comparison and validation, as recommended in the ITU-T Rec. P.564. Due to the relevance of PESQ, this was chosen.

5.1.2 ArQoS®: the call quality monitoring system

In this study, a call quality monitoring system for networks in use at Portugal Telecom Inovação Labs has been used: the ArQoS®. This monitoring system permits to set up, maintain, monitoring and analyse telephony calls over technologies such as PSTN, GSM or IP. It injects a −10 dBm 1 020 Hz signal into the system by which the system where the quality is to be monitored. It is important to state that the use of a 1 020 Hz signal and impedances matching permits to consider the loss of levels as loudness ratings [156, 157].

The ArQoS® is formed by two modules. One module includes a PESQ application. It makes possible the system to provide QoS, such as delays and noise, and QoE metrics, such as MOS. Nevertheless, this module operated as a "black box", since it was developed by Psytechnics, which detains the intellectual property rights. In this manuscript, let us call it the "PESQ module". This application provides some important parameters that are useful in this study. Concerning QoS metrics, it provides:

- Delay values
  - Maximum delay (presented as "*Max delay*"),
  - Average delay (presented as "*Avg delay*") and
  - Minimum delay (presented as "*Min delay*").
- Attenuation (presented as "*Att*"),
- Signal levels
  - Reference level (presented as "*Ref level*") and
  - Degraded level (presented as "*Degraded level*").

Concerning QoE metrics it provides:

- Raw MOS (presented as "*PESQ Score(P.862)*"),
- Subjective MOS (presented as "*MOS P.800*"),
- MOS_{LQO}, (presented as "*MOS LQO (P.862.1)*") and
- E-Model R factor (presented as "*G.107 Rating*").

The other module of ArQoS® was developed in Portugal Telecom Inovação Labs. Let us call it the "PTIn module". It only provides QoS parameters. They are:

- Voice delay,
- Round-trip delay,
- Noise, reduced to the caller side and
- Noise, reduced to the called side.

In this work the ArQoS® system is used to give both the reference MOS value to calibrate the proposed method and to provide QoS values that are used as input parameters.

Since the PESQ module operates as a "black box", there is some degree of uncertainty about the nature of some of these parameters, as is the case of $Max\ delay$, $Avg\ delay$ and $Min\ delay$. These designations suggest absolute values of end-to-end delays. However, since they are provided by the PESQ module but the PESQ algorithm does not take

delays into account (just delay variations, as referred to in section 3.4.1), these temporal values are firstly interpreted as delay variations, as they are provided by the PESQ application. Nevertheless, since this module is a third party application, which may add the features considered as useful, these temporal values can be interpreted as not being relative to PESQ application itself, but simply measured latencies. Similar doubts arise relative to attenuation (*Att*) and signal levels (*Ref level* and *Degraded level*) because they are not part of the PESQ algorithm, too. Does *Att* represent the difference between *Degraded level* and *Ref level*? In case of doubt, all these parameters are considered as being independent.

The adopted methodology involves, on the one hand, first characterising the scenario of interest and identifying all the equipments it encompasses as well as the type of connections that are involved. Then, respective characterising parameter values are gathered and used as input to the E-Model, whether they would be given by ArQoS® or they would be provided by datasheets of equipments. With these values the MOS value is calculated according to the E-Model, as given by expressions 3.4 and 3.5. On the other hand, the PESQ module is run and the resulting MOS values are used as the reference to calibrate the derived model.

The accuracy requirement has been established so that the absolute value of the difference between the calculated MOS and the reference MOS must be less than 0.15. That is:

$$|MOS\ error| < 0.15 \tag{5.1}$$

In the experiences, two models of switches are used as central offices: the Siemens EWSD and the Alcatel System 12.

5.1.3 A preliminary study

In order to get practical insight of this type of approach, a preliminary study was carried out. Whilst ITU-T Rec. G.107 presents the E-Model reference model, as shown in Fig. 3.4, and from which the specific scenarios of interest must be derived, the ITU-T Rec. G.108 specifies two alternative criteria so that practical scenarios can be classified. The first criterion respects to the number of wires in the terminal equipment, regardless

of the type of the central office switches (analog or digital). It can be a 2-wire or a 4-wire equipment. The second criterion respects the remote configuration which can be a private network, a public network or a far-end network, which, in turn, can be divided in European and North American [141].

For this preliminary study, the first criterion is adopted and the 2-wire to 2-wire scenario is chosen. This corresponds to an analog-to-analog end-to-end telephony system. It is shown in Fig 5.1. This figure also shows the input parameters to consider in this scenario. As it can be seen, they are, *grosso modo*, the loudness rates of side A and side B, SLR_A and RLR_B, including the equipment loudness rates, $SLRa$, $RLRa$, $SLRb$ and $RLRb$, the design factors, Ds and Dr, the room noise, Ps and Pr, Weighted Echo Path Loss (WEPL), the mean one-way delay of the echo path, T, the round trip delay in a 4-wire loop, Tr, the absolute delay in echo free connections, Ta, the quantisation distortion unit, qdu, the STMR, the Listener Sidetone Masking Rating (LSMR), Listner Side Tone (LSTR), the TELR and the Advantage factor, A. The meaning of these parameters has been explained in section 3.5.1.

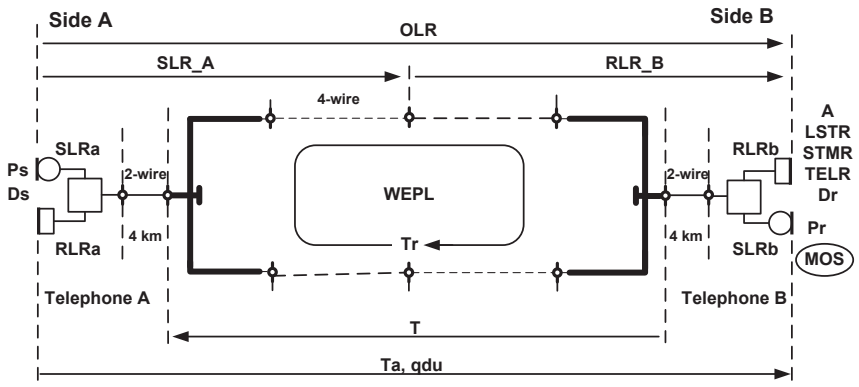


Figure 5.1 – Scenario for 2-wire to 2-wire connections

This scenario is characterised by:

- 2-wire to 2-wire connection, via public network;
- Distance to the central office: 4 km for both subscribers;
- Central office: digital, Siemens EWSD.

Since the considered distances are very low, the considered delay values are considered as negligible:

- T = 0 ms;
- Ta = 0 ms;
- Tr = 0 ms.

According to the recommended procedures of ITU-T Rec. G.108 a "pre-calculation" of $TELR$ and $WEPL$ must be performed:

- $WEPL$ = +110. The default value is used because the value of this parameter is not known for the switch in use (Siemens EWSD).
- $TELR = SLR\_B + RLR\_B + EchoLoss$ = 8 + 2 + 110 = +120 dB.

As *Echo Loss* value, the $WEPL$ value was used, since there are no additional losses in the echo path other than the $WEPL$. All the remaining parameters are set to the default values.

From running the E-Model algorithm as defined in ITU-T Rec. G.107, by applying equations 3.4, 3.5 and 3.6, the values of the "Simultaneous impairment factor", Is, the "Delay impairment factor", Id, the "Equipment impairment factor", Ie, and the "Effective equipment impairment factor", Ie_eff are calculated and so the values of the rating factor, R, and MOS. The obtained values are shown in Table 5.2.

According to Table 3.12, that defines the categories of speech transmission quality as a function of the R value, this value of R (93.21) denotes a voice quality of category "Best" to which corresponds a classification of "Very satisfied" concerning the user satisfaction [142]. The corresponding MOS value is 4.41 according to ITU-T Rec. G.107 (see Eq. 3.5), which is a good value. In fact, the degradations that occur simultaneously with the signal (Is), as well as those caused by delays and echoes (Id), encoders (Ie) or even by packet losses, are practically insignificant in this almost ideal scenario. Such result is in line with the fact that the transmission path is a local path.

The ArQoS® was then run, obtaining the scores shown in Tables 5.3 and 5.4. The former shows the results given by the PESQ module and the second shows the values given by the PTIn module.

Table 5.2 – Preliminary Scores given by the E-Model algorithm

Is	*Id*	*Ie*	*Ie_eff*	*R*	*MOS*
1.413568	0.149046	0.000000	0.000000	93.21	4.41

Table 5.3 – Preliminary Scores given by the PESQ module

PESQ Score (P.862)	3.500
MOS P.800	3.587
MOS_LQO (P.862.1)	3.554
G.107 Rating	69.081
Max delay	10.75 ms
Avg delay	10.704 ms
Min delay	10.625 ms
Att	+6.891 dBov
Ref level	−34.426 dBo
Degraded level	−46.607 dBov

Table 5.4 – Preliminary Scores given by the ArQoS® PTIn module

Voice delay	7 ms
Round-trip delay	14 ms
Noise, caller side	−74.565 dB
Noise, called side	−76.905 dB

As it can be seen by comparing the MOS values presented either in Table 5.2 and Table 5.3, there is a discrepancy between the calculated result (4.41) and those obtained from the PESQ module (3.500, 3.587 and 3.554). It figures out that additional experiences have to be carried out in order to refine the method and thus improve MOS accuracy.

Furthermore, another discrepancy is found in the temporal values when comparing results of Table 5.3 with those of Table 5.4. For sure, these parameters do not represent the same reality. It also figures out that additional experiences might clarify the role of each of these QoS parameters. Another interpretation concerning the cause of such divergences is that too many default values are being used in the input parameters. Thus, for the

definitive calculations regarding the scenario of interest, it is necessary to know the characteristic values of the switch in use, specifically the values of *WEPL*, *TELR* and the internal attenuation in order to infer the values of *SLR* and *RLR*.

Taking into account the previous discussion, those parameters that contribute for better accuracy and also a maximum number of known values to be used as input in the calculations have to be identified. For this purpose, the QoS values given by ArQoS® are used. They essentially include attenuations, delays and noise values. Table 5.5 identifies the available parameters and their origin. Input values, like *WEPL* and *TELR* are derived from these parameters.

Table 5.5 – Input parameters which values are available from measuring

Type of parameters	**Parameters**	**Origin**
Attenuation	*Att*	PESQ module
	Ref level	
	Degraded level	
Delay	*Max delay*	
	Avg delay	
	Min delay	
	Round-trip delay	PTIn module
	Voice delay	
Noise	*Noise, caller side*	
	Noise, called side	

However, since there is no obvious correspondence between the available ArQoS® parameters shown in Table 5.5 and the input parameters to be used in the E-Model, it is necessary to determine the following:

- Which attenuation parameter best represents the needed loudness rating;
- Which ArQoS® delay parameter best represent the E-Model delay parameter(s);
- Which ArQoS® noise parameter best represents the noise parameters of the E-Model.

To find these parameters, some assumptions must be made. For instance, by assuming that *Att* corresponds to the end-to-end attenuation and so, the OLR, or, alternatively, that the difference between *Ref level* and *Degraded level* can also provide the OLR. Therefore, it is possible to realise that different measures can be used as the origin of the OLR

parameter. Furthermore, for each of these origins, there are several variants to consider, such as to include or not the terminal equipments (telephones) in the OLR calculation. To sum up, combining all possible origins of the parameters with all variants, multiple cases may be considered to provide the OLR values. The same kind of assumptions can be made in regard to the remaining parameters such as the delay and the noise.

Thereafter, by combining all the OLR cases with all the delay, noise, *WEPL* and *TELR* cases, a set of patterns of input parameters (*i.e.*, the resulting combinations) are formulated and used to calculate the MOS. An example of a pattern of input parameters is the following:

- *Att* as the Overall Loudness Rate, OLR, and not include telephones;
- *Noise, caller side* as the circuit noise referred to the point 0 dBr, Nc;
- *Avg delay* as the absolute delay in echo free connections, Ta;
- *WEPL* as the minimum recommended value;
- return losses in the hybrid, Lr as the typical value;
- *TELR* as the sum of SLR, RLR, $WEPL$ and Lr.

Following this kind of reasoning, several different patterns can be identified. The aim is then to find which of them leads to a calculated MOS that best approximates the reference (measured) MOS.

The described methodology is used on both local calling area and far end calling scenarios. Section 5.2 presents the study that permits to derive the evaluation module relative to the former and section 5.3 presents the study that permits to derive the module relative to the far-end calling scenario.

5.2 The local calling area module

The study presented in this section comprises, firstly, gathering the input parameters by identifying the cases that may be part of the input patterns and, secondly, the experimental results and subsequent discussion.

5.2.1 Gathering of input parameters

In this subsection the possible cases that can provide values for the loudness rate, noise, delays, *WEPL* and *TELR*, that fit this scenario, are identified.

Cases concerning loudness rates

Concerning the attenuation effect, the identified cases are:

- **H0**: Use the default values:
  - $SLR = 8$; $RLR = 2$; $OLR = 10$.
- **H1.1**: Use the *Att* value as the end-to-end attenuation. This is assumed to be the OLR value. In this case a proportional adjustment is done taking into account the default proportionality between SLR and RLR, that is, 8/10 and 2/10, respectively. This is previously referred to as origin of the value.
- **H1.2**: The same origin and values as in H1.1 are used, except the default values of SLR and RLR, which are added in this case. The reason for this variant is that, according PTIn staff, the PESQ application injects the original voice signal into the Pulse Code Modulation (PCM) bus (not into the telephone handset). Hence, to simulate the complete pathway including the passage of the signal through the telephones, the respective loudness rates, taken into account in parameters *SLRa*, *RLRa*, *SLRb* and *RLRb* must be added.
- **H2.1**: A different origin for OLR is identified: the value of OLR is obtained by the difference between the *Ref level* and *Degraded level* values. The same proportionality adjustment as in H1.1, and for the same reasons, is done in this case.
- **H2.2**: The same origin and values used in H2.1 are defined but with the variant in which the default values of SLR and RLR are added, as in H1.2, for the same reasons.
- **H3.1**: The values given by the PTIn module are used as the origin of the OLR. To derive the OLR value it was considered that the 1 020 Hz signal of −10 dBm0 is injected either in the calling side (send side) and measured at the receive side, either injected in the receive side and measured in the send side. Thus the calculated OLR

results from the average of these differences of level, as follows:

$$OLR = \frac{(-10 - receive\ side\ level) + (-10 - send\ side\ level)}{2}\ [dB]. \tag{5.2}$$

The SLR and RLR values were again derived by applying the proportionality mentioned above, as described for H2.1.

- **H3.2**: The same origin and the same values as in H3.1 are used, except that the default values of SLR and RLR are now added for the same reasons as mentioned in the cases H1.2 and H2.2.
- **H3.3**: The 1 020 Hz signal of −10 dBm0 is considered to be injected in an intermediate point of the circuit −the 0 dBr point as defined in ITU-T Rec. G.106. In this case the difference between the level of this signal and the level measured at the send side is interpreted as the SLR value. For this reason, the attenuation occurred in the direction caller to called part is assumed to be equal to the attenuation occurred in the reverse direction in the same path.

$$SLR = -10 - send\ side\ level\ [dB]. \tag{5.3}$$

Similarly, the difference between −10 dBm0 and the level that is measured at the receive side is interpreted as being the RLR value. For this, it is also assumed that the attenuation occurred in the direction send side to receive side is equal to the attenuation occurred in the reverse direction (receive side to send side).

$$RLR = -10 - receive\ side\ level\ [dB]. \tag{5.4}$$

- **H3.4**: The same values as in H3.3 are used and the default values of SLR and RLR are added for the same reasons as pointed out in H1.2, H2.2 and H3.2.

Cases concerning noise parameters

The noise value used in the E-Model must be referred to the 0 dBr point. Since the noise values available from the measurements come from both the send and receive sides, it is necessary to reduce them to the 0 dBr point. This is done by subtracting them the attenuation values (loudness rating values) corresponding to the path between each side

(either send or receive side) and the 0 dBr point [141]. That is,

$$Nc = Measured\ level\ at\ a\ side - LR, \tag{5.5}$$

where LR represents either SLR or RLR.

As it can be seen, all the calculated Nc values will depend on the previously identified attenuation cases since they depend on the LRs that have been derived.

To take into consideration the noise effects, the next cases are identified:

- **I0**: The default noise value, Nc, is used: $Nc = -70$ dBm.
- **I1.1**: As origin of the Nc value, uses the noise measured at the send side. As a variant, the loudness rate calculated in the case H1.1 is used to reduce the value to the 0 dBr point.
- **I1.2**: Uses the same origin as in I1.1 (send side) and the loudness rate value calculated in the case H2.1 to reduce the noise value to the 0 dBr point.
- **I1.3**: Uses the same origin as in the case I1.1 and the loudness rate value calculated in the case H3.1.
- **I1.4**: Uses the same origin as in I1.1 and the loudness rate value calculated in the case H3.3.
- **I2.1**: As origin, uses the noise measured at the receive side and the variant obtained from using the loudness rate value calculated in the case H1.1.
- **I2.2**: Uses the same origin as in I2.1 (receive side) and the loudness rate value calculated in the case H2.1.
- **I2.3**: Uses the same origin as in I2.1 and the loudness rate value calculated in the case H3.1.
- **I2.4**: Uses the same origin as in I2.1 and the loudness rate value calculated in the case H3.3.

Notice that the values considered in cases H1.2, H2.2 and H3.3 are not considered in the reduction operation because it is known that noise is not measured at the telephone handsets, since the laboratory setup does not include the handsets.

Cases concerning delays

To take into consideration the delay effects, the next cases are identified:

- **J0**: The default values are used: T = 0 ms; Ta = 0 ms and Tr = 0 ms;
- **J1**: The *Avg delay* is used;
- **J2**: The *Voice delay* is used.

Cases concerning *WEPL*

To find which is the best *WEPL* value for the switch, three cases are identified: i) the minimum, ii) the maximum (which coincides with the default value) of the permitted range and iii) zero for the purposes defined below.

- **L1**: *WEPL*= 0 dB;
- **L2**: *WEPL*= +5 dB (minimum value);
- **L3**: *WEPL*= +110 dB (default value).

Cases concerning returning losses in the hybrid

Since the returning losses in the hybrid, Lr, also contribute to the *WEPL* value, two cases are identified:

- **M1**: Lr = 0 dB. This value is necessary to match with the cases L2 and L3;
- **M2**: Lr = +17 dB. This is a typical value to this parameter as referred in [158]. It might match the case L1.

Cases concerning *TELR*

The pre-calculation of the $TELR$ values is given by the expression 5.6. The values derived by Eq. 5.6 constitute the derived cases concerning $TELR$ to integrate, latter, the previously referred input patterns. The default value ($TELR$ = 65) is also added to this set of cases.

$$TELR = SLR + RLR + WEPL + Lr \ [dB]. \tag{5.6}$$

Lr represents the return losses in the hybrid, which is also part of the echo path [158].

In order to carry out the simulation tests, a program in Matlab® was developed. It combines all the identified cases concerning the analysed parameters to constitute all the patterns of inputs and, for each one, calculates a MOS value.

5.2.2 Results and discussion

Based on the methodology previously described, the MOS values relative to all input patterns were calculated and compared with the reference MOS obtained from running the PESQ module. From this comparison, the input pattern that best approximates the calculated MOS to the measured MOS is identified. Thus, the cases that constitute each pattern can be identified as the best inputs for the derived model.

The MOS errors (the differences between the calculated MOS and that given by PESQ) obtained from the previous methodology range from −2.64 to 1.04. By restricting all the patterns found through the process described above to those in which the calculated MOS values fall into the error bound specified by Eq. 5.1, it is possible to select the patterns that lead to relatively accurate MOS values. Fig 5.2 and 5.3 show the MOS error for the best patterns. The x-axis represents the numbers given to each identified pattern. They are sorted by the MOS error value. The y-axis represents the respective MOS error. Vertical lines delimit the patterns for which the MOS errors are similar. Representation of the patterns is restricted to those that led to the best results. As it can be seen, each of the different switches lead to different results.

Fig 5.2 shows the MOS errors relative to the Siemens EWSD switch. From the results obtained under these test conditions, it is possible to verify that the lower errors, plotted in the centre of the figure, comprise the cases marked with the labels H3.4, L3 and J0, J1, J2. Referring to the previously identified cases, it denotes that the relevant factors are essentially attenuation (case H3.4) and the echo (L3). It is also possible to verify that the cases that include the sum of the default values of SLR and RLR (case H3.4) are those that lead to the more accurate results. Timing parameters are not so relevant and noise parameters revealed the less relevant factors for the conditions of test. For these cases, the calculated errors verify the condition $|MOS\ error| < 0.09$. By convenience let us designate H3.4 and L3 as the *suitable cases* of the Siemens switch.

Fig. 5.3 shows the MOS errors relative to the Alcatel System 12. As it can be seen, the combination of the cases H1.2, L1 and J0 are those that lead to the lower errors. Calculated values are such that $|MOS\ error| < 0.014$. The remaining cases do not influence the results for the same test conditions. Let us designate H3.3, L1 and J0 as the *suitable cases* of the Alcatel switch.

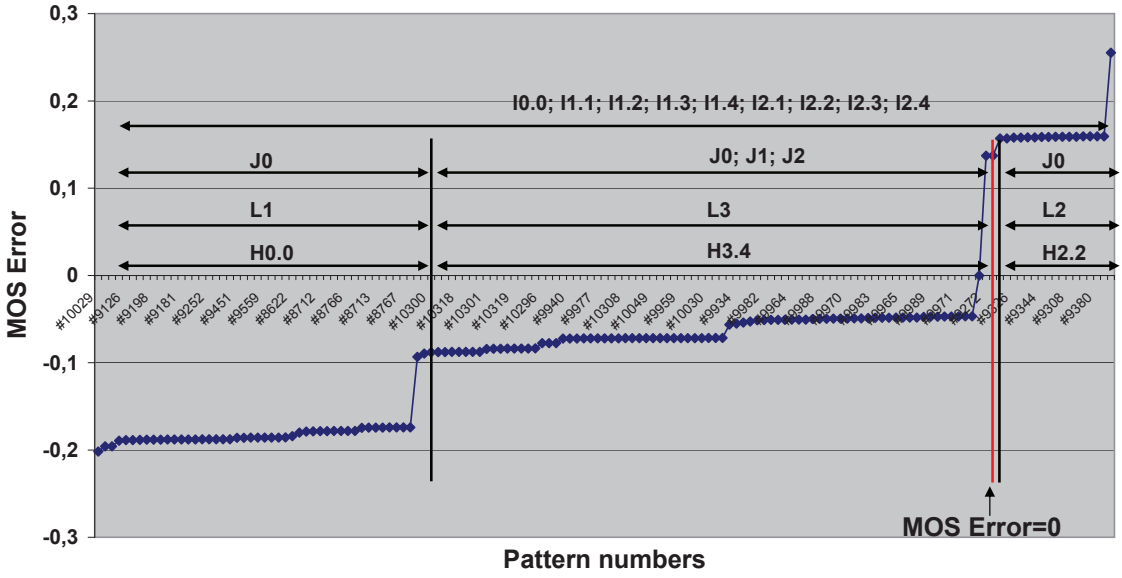


Figure 5.2 – MOS errors obtained for each pattern of inputs (Siemens switch)

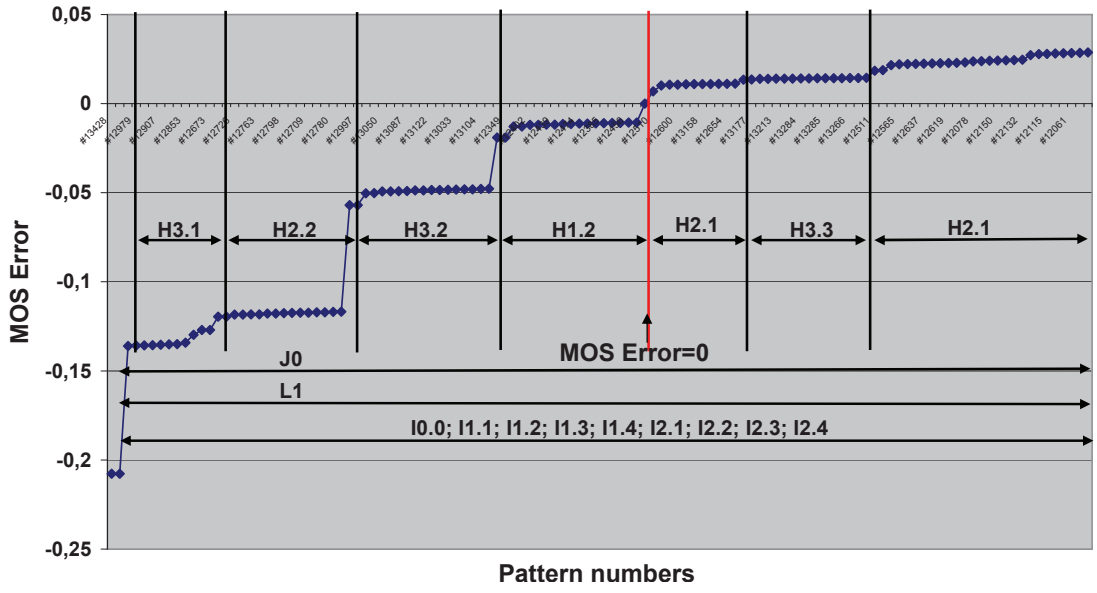


Figure 5.3 – MOS errors obtained for each pattern of inputs (Alcatel switch)

Since the suitable cases for the Siemens switch are different from those of the Alcatel switch, the errors that occur from applying the Siemens suitable cases to the Alcatel switch and *vice-versa* are analysed for the case of using the same cases as inputs of both switch models. The results of this analysis are shown in Table 5.6.

Table 5.6 – $|MOS\ Error|$ that results from applying the same cases to both switches.

	Siemens suitable cases	**Alcatel suitable cases**
Siemens switch	0.09	0.29
Alcatel switch	0.51	0.014

As it can be observed in Table 5.6, the Alcatel suitable cases are those that give the best balanced results in the case where the same patterns are used in both switches: 0.014 if used in Alcatel switches; 0.29 if used in Siemens switches. Otherwise, by applying the Siemens suitable cases to the Alcatel switch, a minimum error of 0.51 is achieved. Suitable cases of a switch are hereinafter called *foreign cases* when applied to the other switch.

As mentioned before, these results were obtained for a communication scenario where a unique switch is included in the transmission path. Next section is concerned with a scenario in which two switches are included in the pathway and connected via SS7 protocol[1]. This scenario simulates the public, even the far-end network, as defined in the E-Model in which a switch is placed at each end side.

5.3 The long distance calling module

This scenario is defined to study the case where two switches are connected by the SS7 signalling system in order to simulate a more comprehensive geographical area. Fig 5.4 depicts the interconnection between two switches: Siemens EWSD and Alcatel System 12. In order to calculate the OLR of the entire path given by the scenario of Fig. 5.4, it is important to determine if OLR can be calculated by simply summing the individual OLR from each switch contribution. Thus, some preliminary calculations are performed.

[1]SS7 stands for Signalling System No.7, which is in fact a protocol suite that is used to set up and tear down most of the world's public switched telephone network telephone calls. It implements the relatively well known Channel Associated Signaling (CAS) and the Common Channel Signaling (CCS) that are present in primary PCM systems as the 1.544 Mbps American T1 trunk and 2.048 Mbps European E1 trunk, respectively.

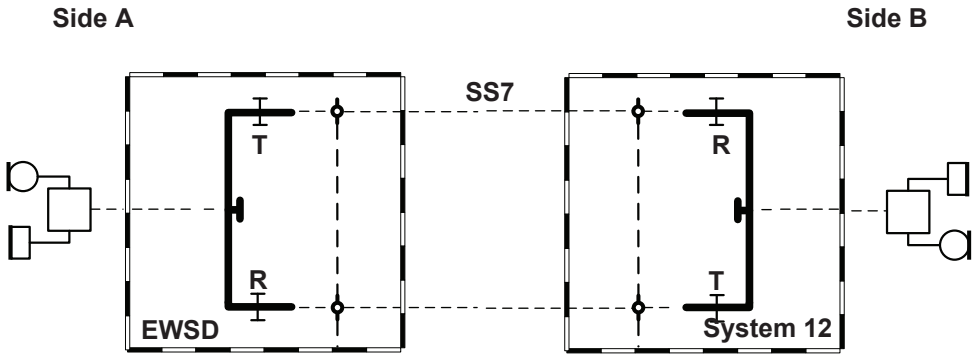


Figure 5.4 – Interconnection between two switches: Siemens EWSD and Alcatel System 12

5.3.1 Scenario updating and preliminary calculations

To determine if the OLR results from adding individual OLRs, the individual attenuations as well as their sum were calculated for the most significant cases, as determined in section 5.2. Table 5.7 shows these individual attenuations and their sum, for the most significant identified cases. Note that these attenuations were measured by using the ArQoS® system. The respective values, for the same significant cases are shown in Table 5.8.

Table 5.7 – Calculated attenuation [dB] for the scenario Siemens EWSD + Alcatel System 12

	Most significant cases				
Scenario	**H1.1**	**H2.1**	**H3.1**	**H3.3**	**H3.4**
Siemens	6.98	12.29	8.88	17.77	27.77
Alcatel	4.12	9.22	6.18	12.37	22.37
Siemens + Alcatel	11.07	21.56	15.05	30.12	50.12

Table 5.8 – Measured attenuation [dB] for the scenario Siemens EWSD + Alcatel System 12

	Most significant cases				
Scenario	**H1.1**	**H2.1**	**H3.1**	**H3.3**	**H3.4**
Siemens and Alcatel connected	7.67	13.81	8.80	17.60	27.60

As it can be seen, taking the example given by the case H1.1, the sum of the calculated attenuations (11.07 dB) is not the same as the measured attenuation (7.67 dB). From

the obtained values it is possible to conclude that the overall attenuation cannot be calculated by simply summing the individual attenuations of each switch, as calculated in subsection 5.2.2. A partial interpretation is that the whole attenuation is the result of partial contributions from both switches, which value is unknown.

In order to better characterise the problem of calculating the OLR of the scenario represented in Fig. 5.4, it was considered necessary to refine the E-Model implementation such that more details about attenuation can be included in the model.

As previously mentioned, the E-Model reference divides the end-to-end path into two sides: send side (side A) and receive side (side B). The intermediate point, designed by 0 dBr point, establishes two sides whose attenuations for a 1 020 Hz signal are defined as "Send Loudness Rating" (*SLR*), concerning the send side and the "Receive Loudness Rate" *RLR*, concerning the receive side, when impedances match. According to the definitions given in section 5.1.3, let us call them SLR_A and RLR_B, respectively. Similarly, SRL_B and RLR_A stand for the Send Loudness Rating and Receiver Loudness Rate when considering the reverse direction, *i.e.*, from side B to side A. The need to consider such reciprocity is based on the hypothesis that attenuation in one direction may be different from attenuation in the reverse direction.

In order to better guarantee an accurate implementation of the E-Model let us examine in detail the parameters that form SLR_A and RLR_B. Fig. 5.5 shows the relevant parameters to use, for which the definitions are given below.

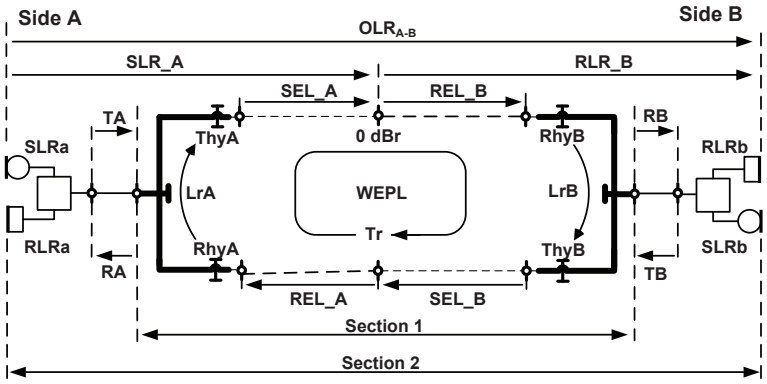


Figure 5.5 – Detailed scenario for the E-Model implementation refining

The parameters represented in Fig. 5.5 stand for:

- *SLRa*: losses occurred in the telephone of side A when transmission is done from side A to side B;
- *TA*: losses occurred in the subscriber line of side A when transmission is done from side A to side B;
- *ThyA*: losses occurred in the hybrid of side A, in the 2-wire to 4-wire transition when transmission is done from side A to side B;
- *LrA*: return losses occurred in the hybrid of the side A;
- *SEL_A* (Send Echo Loss on side A): losses occurred in the echo circuit of side A, when transmission is done from side A to side B;
- *REL_B* (Receive Echo Loss on side B): losses occurred in the echo circuit in side B, when transmission is done from side A to side B;
- *RhyB*: losses occurred in the hybrid of side B, in the 4-wire to 2-wire transition, when transmission is done from side A to side B;
- *RB*: losses occurred in the subscriber line of side B when transmission is done from side A to side B;
- *RLRb*: losses occurred in the telephone of side B when transmission is done from side A to side B;
- *SLRb*: losses occurred in the telephone of side B when transmission is done from side B to side A;
- *TB*: losses occurred in the subscriber line of side B when transmission is done from side B to side A;
- *ThyB*: losses occurred in the hybrid of side B, in the 2-wire to 4-wire transition when transmission is done from side B to side A;
- *LrB*: return losses occurred in the hybrid of the side B;
- *SEL_B*: losses occurred in the echo circuit in side B, when transmission is done from side side B to side A;
- *REL_A*: losses occurred in the echo circuit in the side A when transmission is done

from side B to side A;

- $RhyA$: losses occurred in the hybrid of side A, in the 4-wire to 2-wire transition when transmission is done from side B to side A;
- RA: losses occurred in the subscriber line of side A when transmission is done from side B to side A;
- $RLRa$: losses occurred in the telephone of side A when transmission is done from side B to side A.

Now, using the parameters defined above, the necessary pre-calculations are obtained as follows:

- Talker Echo Loudness Rate ($TELR$), experienced by a speaker that is placed in the side B:

$$\begin{aligned} TELR|_B &= SLRb + TB + ThyB + SEL\_B + REL\_A + \\ &\quad + RhyA + LrA + ThyA + SEL\_A + \\ &\quad + RLR\_B + RhyB + RB + RLRb; \end{aligned}$$

- Weighted Echo Path Loss ($WEPL$):

$$\begin{aligned} WEPL &= REL\_A + RhyA + LrA + ThyA + SEL\_A + REL\_B + \\ &\quad + RhyB + LrB + ThyB + SEL\_B; \end{aligned} \tag{5.7}$$

- Send Loudness Rate of side A:

$$SLR\_A = SLRa + TA + ThyA + SEL\_A; \tag{5.8}$$

- Receive Loudness Rate of side B:

$$RLR\_B = RLR\_B + RhyB + RB + RLRb; \tag{5.9}$$

- Overall Loudness Rate in the direction A to B:

$$\begin{aligned} OLR_{A\to B} &= SLRa + TA + ThyA + SEL\_A + REL\_B + \\ &\quad + RhyB + RB + RLRb \\ &= SLR\_A + RLR\_B. \end{aligned}$$

- And reciprocally,

$$OLR_{B\to A} = SLR\_B + RLR\_A.$$

- Noise measured in side B, referred to the 0 dBr point:

$$Nc = measured\ noise - TB - ThyB - SEL\_B. \tag{5.10}$$

Since there is some redundancy in the described parameters, it is possible to achieve some simplification by assuming some conditions:

- Null losses in the subscriber line, since in the experiences this length is virtually zero.

  $TA = TB = RA = RB = 0$ dB;

- Null losses in the 4-wire section, for the same reason.

  $SEL\_A = SEL\_B = REL\_A = REL\_B = 0$ dB;

- Symmetry in the values of the following attenuations:

  $SLRa = SLRb$;

  $RLRa = RLRb$;

  $ThyA = ThyB = RhyA = RhyB$;

  LrA=LrB.

  This symmetry permit us to consider that

  $OLR_{A\rightarrow B} = OLR_{B\rightarrow A}$.

5.3.2 Gathering of input parameters

Based on the previous considerations, a new set of cases is formulated, applied to the current scenario.

Cases concerning loudness rating

Concerning loudness rating parameters, next cases are considered. Due to practical feasibility reasons, only the values given by the ArQoS® PTIn Module are considered.

- **A00**: Considers that attenuations are exclusively due to the terminal equipments (telephones) and that they are analog, as defined in [159]:

$SLRa$ = +11 dB;

$RLRb$ = −3 dB.

- **A01**: Considers that attenuations are exclusively due to the terminal equipments. Uses the default values for the digital telephones, according to [141]:

  $SLRa$ = +7 dB;

  $RLRb$ = +3 dB.

- **A11**: Considers that the reference signal of −10 dBm 1 020 Hz is injected in the 0 dBr point. The difference between the reference signal level and the level measured in the origin (calling side, side A) is considered as being SLR_A. Similarly, RLR_B results from the difference between the reference signal and the level measured in the called party (side B). Section 1 of the Fig. 5.5 is being considered, thus not including the Loudness Rate values relative to the terminal equipments.

  $SRL\_A = -10 - side\ A\ level$ [dB];

  $RLR\_B = -10 - side\ B\ level$ [dB];

  $OLR = SLR\_A + RLR\_B$ [dB].

- **A12**: Similar to A11 except that section 2 of Fig. 5.5 is now considered. Equipments are analog, as in the case A00.

  $SLR\_A = (-10 - side\ A\ level) + SLRa = (-10 - side\ A\ level) + 11$ [dB];

  $RLR\_B = (-10 - side\ B\ level) + RLRb = (-10 - side\ B\ level) - 3$ [dB];

  $OLR = SLR\_A + RLR\_B$ [dB].

- **A13**: The reference signal is considered as injected in the calling part (side A) and section 1 considered, too. Thus the measured difference of levels corresponds to $OLR_{A\to B} = SLR\_A + RLR\_B$. In order to derive SLR_A and RLR_B, the 0 dBr point is considered as the point in the circuit where the attenuation is half of the whole attenuation.

  $OLR_{A\to B} = -10 - side\ B\ level$ [dB];

  $SLR\_A = 0.5 \times OLR_{A\to B}$ [dB];

  $RLR\_B = 0.5 \times OLR_{A\to B}$ [dB].

- **A14**: Similar to A13, except that the 0 dBr point results from the proportionality

SLR_A/RLR_B = 8/2, based on the proportionality of the SLR/RLR that is 8/2, too.

$OLR_{A\rightarrow B}$ = −10 − *side B level* [dB];

$SLR\_A = 0.8 \times OLR_{A\rightarrow B}$ [dB];

$RLR\_B = 0.2 \times OLR_{A\rightarrow B}$ [dB].

- **A15**: Similar to A13, except that the section 2 (Fig. 5.5) is, now, considered and the end terminals are considered as analog.

$$\begin{aligned} OLR_{A\rightarrow B} &= (-10 - side\ B\ level) + SLRa + RLRb \\ &= (-10 - side\ B\ level) + 11 - 3\ [dB]; \end{aligned}$$

$SLR\_A = 0.5 \times OLR_{A\rightarrow B}$ [dB];

$RLR\_B = 0.5 \times OLR_{A\rightarrow B}$ [dB].

- **A16**: Similar to A15 for which a proportionality of 8/2 is considered.

$$\begin{aligned} OLR_{A\rightarrow B} &= (-10 - side\ B\ level) + SLRa + RLRb \\ &= (-10 - side\ B\ level) + 11 - 3\ [dB]; \end{aligned}$$

$SLR\_A = 0.8 \times OLR_{A\rightarrow B}$ [dB];

$RLR\_B = 0.2 \times OLR_{A\rightarrow B}$ [dB].

- **A17**: Similar to A13 for which the reference signal is considered injected in the called party (side B).

$OLR_{B\rightarrow A}$ = −10 − *side A level* [dB];

$SLR\_A = 0.5 \times OLR_{B\rightarrow A}$ [dB];

$RLR\_B = 0.5 \times OLR_{B\rightarrow A}$ [dB].

- **A18**: Similar to A17 in which the considered proportionality is 8/2.

$OLR_{B\rightarrow A}$ = −10 − *side A level* [dB];

$SLR\_A = 0.8 \times OLR_{B\rightarrow A}$ [dB];

$RLR\_B = 0.2 \times OLR_{B\rightarrow A}$ [dB].

- **A19**: Similar to A17 in which the section 2 (Fig. 5.5) is, now, considered and the end terminals are considered analog.

$$\begin{aligned} OLR_{B\rightarrow A} &= (-10 - side\ A\ level) + SLRb + RLRa \\ &= (-10 - side\ A\ level) + 11 - 3\ [dB]; \end{aligned}$$

$SLR\_A = 0.5 \times OLR_{B\rightarrow A}$ [dB];

$RLR\_B = 0.5 \times OLR_{B\to A}$ [dB].

- **A20**: Similar to A19 in which the proportionality is, now, 8/2.

$$\begin{aligned} OLR_{B\to A} &= (-10 - side\ A\ level) + SLRb + RLRa \\ &= (-10 - side\ A\ level) + 11 - 3\ [dB]; \end{aligned}$$

$SLR\_A = 0.8 \times OLR_{B\to A}$ [dB];

$RLR\_B = 0.2 \times OLR_{B\to A}$ [dB].

- **A21**: Similar to A12 in which the default values concerning $SLRa$ and $RLRb$ refer, now, to digital telephones.

$SLR\_A = (-10 - side\ A\ level) + SLRa = (-10 - side\ A\ level) + 7$ [dB];

$RLR\_B = (-10 - side\ B\ level) + RLRb = (-10 - side\ B\ level) + 3$ [dB];

$OLR_{A\to B} = SLR\_A + RLR\_B$ [dB].

- **A22**: Similar to A15 in which the default values concerning $SLRa$ and $RLRb$ refer, now, to digital telephones.

$$\begin{aligned} OLR_{A\to B} &= (-10 - side\ B\ level) + SLRa + RLRb \\ &= (-10 - side\ B\ level) + 7 + 3\ [dB]; \end{aligned}$$

$SLR\_A = 0.5 \times OLR_{A\to B}$ [dB];

$RLR\_B = 0.5 \times OLR_{A\to B}$ [dB].

- **A23**: Similar to A15 in which the default values concerning $SLRa$ and $RLRb$ refer now to digital telephones.

$$\begin{aligned} OLR_{A\to B} &= (-10 - side\ B\ level) + SLRa + RLRb \\ &= (-10 - side\ B\ level) + 7 + 3\ [dB]; \end{aligned}$$

$SLR\_A = 0.8 \times OLR_{A\to B}$ [dB];

$RLR\_B = 0.2 \times OLR_{A\to B}$ [dB].

- **A24**: Similar to A22 in which the reference signal is considered as injected in the called party (side B).

$$\begin{aligned} OLR_{B\to A} &= (-10 - side\ A\ level) + SLRb + RLRa \\ &= (-10 - side\ A\ level) + 7 + 3\ [dB]; \end{aligned}$$

$SLR\_A = 0.5 \times OLR_{B\to A}$ [dB];

$RLR\_B = 0.5 \times OLR_{B\to A}$ [dB].

- **A25**: Similar to A24 in which the considered proportionality SLR_A/RLR_B is, now, 8/2.

$$
\begin{aligned}
OLR_{B\to A} &= (-10 - side\ A\ level) + SLRb + RLRa \\
&= (-10 - side\ A\ level) + 7 + 3\ [dB];
\end{aligned}
$$

$$SLR\_A = 0.8 \times OLR_{B\to A}\ [\text{dB}];$$

$$RLR\_B = 0.2 \times OLR_{B\to A}\ [\text{dB}].$$

Cases concerning *WEPL*

Based on the cases characterised by the attenuation, new cases concerning the Weighted Echo Path Loss ($WEPL$) parameter are formulated. Since the calculation of WEPL values is based on loudness rate values let us call "WEPL cases associated to attenuation cases". To identify them, Table 5.9 has been built. It shows these cases and their associated, when applicable. All values rely on the attenuation cases, excepted the default, minimum and maximum values.

Table 5.9 – Cases concerning the *WEPL* value

***WEPL* Cases**	**Associated cases**	**Remarks**
W00	n.a.	$WEPL = 0$.
W01	n.a.	Minimum WEPL value according to [141]. $WEPL = +5$ [dB].
W02	n.a.	Maximum WEPL value according to [141]. $WEPL = +110$ [dB]. Corresponds to the default value.
WA00	A00	*WEPL=SLRa+RLRb+17+17+SLRb+RLRa*
WA01	A01	
W11	A11	*WEPL=SLR_A+17+RLR_B+17+SLR_B+RLR_A*
W12	A12	
W13	A13	
W14	A14	
W15	A15	
W16	A16	
W17	A17	
W18	A18	
W19	A19	
W20	A20	
W21	A21	
W22	A22	
W23	A23	
W24	A24	
W25	A25	

Cases concerning *TELR*

Similarly, cases concerning the (*TELR*) parameter are formulated. As in the case of the *WEPL*, *TELR* values are based on the attenuation values. Thus, the *TELR* cases have their associated attenuation cases. Table 5.10 shows the identified cases.

Table 5.10 – Cases concerning the *TELR* value

***TELR* cases**	**Associated cases**	**Remarks**
T00	n.a.	Minimum TELR value according to [141]. $TELR$ = +5 [dB].
T01	n.a.	Maximum TELR value according to [141]. $TELR$ = +65 [dB]. Corresponds to the default value.
TA00	A00	*TELR=SLRa+RLRb+17+SLRb+RLRa*
TA01	A01	
T11	A11	*TELR=SLR_A+RLR_B+17+SLR_B+RLR_A*
T12	A12	
T13	A13	
T14	A14	
T15	A15	
T16	A16	
T17	A17	
T18	A18	
T19	A19	
T20	A20	
T21	A21	
T22	A22	
T23	A23	
T24	A24	
T25	A25	

Cases concerning noise

To obtain the noise value referred to the 0 dBr point, the average of both values obtained in the calling party (side A) and in the called party (side B) was used, after the reduction referred in Eq. (5.10), that is,

$$Nc = \frac{noise\ measured\ at\ side\ A - SLR\_A + noise\ measured\ at\ side\ B - SLR\_B}{2}. \quad (5.11)$$

Based on expression 5.11 and the attenuation cases, the cases concerning the noise calculation are formulated and showed in Table 5.11.

Table 5.11 – Cases concerning the *Nc* value

***Nc* cases**	**Associated cases**	**Remarks**
N00	n.a.	Default Nc value according to [141]. $Nc = -70$ [dB].
NA00	A00	*Nc=(side A measured noise-SLRa+side B measured noise-SLRb)/2*
NA01	A01	
N11	A11	*Nc=(side A measured noise-SLR_A+side B measured noise-SLR_B)/2*
N12	A12	
N13	A13	
N14	A14	
N15	A15	
N16	A16	
N17	A17	
N18	A18	
N19	A19	
N20	A20	
N21	A21	
N22	A22	
N23	A23	
N24	A24	
N25	A25	

Cases concerning delay

Concerning the delay parameters, T, Ta and Tr, the values coming from the situations referred in the column remarks of the Table 5.12 were used.

Table 5.12 – Cases concerning the *T*, *Ta* and *Tr* values

T, Ta and Tr Associated cases	**Cases base**	**Remarks**
D00	n.a.	Default values according to [141]. $T = Ta = 0$. $Tr = 2 \times Ta = 0$ [dB].
D11	n.a.	$T = Ta = ArQoS\ value$. $Tr = 2 \times Ta$ [dB].

In the next section, the values obtained are presented and discussed.

5.3.3 Results and discussion

In the previous subsection, cases relative to attenuation, noise, *WEPL*, *TELR* and delays have been identified in order to formulate a set of input patterns to feed the E-Model and to identify which of them lead to lower MOS error, using a similar procedure as done in section 5.2, concerning the local calling area scenario. The same strategy of comparison calculated MOS with measured MOS is now being done.

Siemens EWSD

Concerning the Siemens switch, MOS errors range from −2.6103 to +0.8562 which means absolute MOS errors falling into the interval [0, 2.6103]. By constraining the patterns to those for which the error is less than 0.01, the patterns of Table 5.13 are identified as leading to the lower errors. Table 5.13 is sorted by $|MOS\ Error|$ parameter (3rd column).

Table 5.13 – MOS errors obtained for each of the patterns for the Siemens switch

Input patterns		|MOS Error|	Remarks
#	Combination of cases		
1	A21; N21; D00; W21; T21	0.050	The best pattern that use measured values, except delay values.
2	A21; N00; D00; W02; T01	0.054	Remaining default values.
3	A21; N21; D11; W21; T21	0.055	The best pattern using measured values.
4	A12; N00; D00; W01; T01	0.058	Predominance of default values.
5	A12; N00; D11; W02; T01	0.062	Predominance of default values, except delay values.
6	A21; N00; D11; W02; T01	0.067	Predominance of default values, except delay values.
7	A12; N00; D00; W02; T01	0.073	
8	A12; N12; D11; W12; T12	0.079	Second best pattern that uses measured values.
9	A12; N12; D00; W12; T12	0.089	Second best pattern that uses measured values, except delay values.

As it can be seen, the best pattern involves the attenuation cases A21 and its associated cases, N21, W21 and T21, regardless of whether the default delay value ($T = 0$, case D00, row 1) is used or the measured value is used (case D11, row 3). In these patterns the maximum absolute value of the MOS error is $|MOS\ error| < 0.055$. (See row 1 and row 3 of the Table 5.13).

By examining the row 2 it is also possible to state that the use of the default values relative to the Talker Echo Loudness Rate (case T01; $TELR$=65), Weighted Echo Path Loss (case W02; $WEPL$=110), Noise (case N00, Nc=-70 dB) and delay (case D00; T=0) in conjunction with case A21, leads to good results since the absolute MOS error value is $|MOS\ error|$ = 0.054.

The second better pattern that still uses the measured values is shown in row 8. It includes the case A12 and its associated N12, W12 and T12 cases. In this pattern, the absolute MOS error value is $|MOS\ error|$ = 0.079. In the case of using the default delay value (T=0 ms) the absolute MOS error value is $|MOS\ error|$ = 0.089 (row 9 of Table 5.13). These values are so close because the measured delay is not too much different since the used circuit is very short. In general, pattern of row 8 is preferable, since it uses real (measured) values.

The conjunction of this case (A12) with the default values of $WEPL$, $TELR$, Nc and T leads to an absolute MOS error, $|MOS\ error|$ = 0.073. (See row 7 of the Table 5.13).

Table 5.13 shows the best achieved results, including the ones referred to above. Notice that all MOS error values are $|MOS\ error| < 0.1$, which matches the requirement specified in expression 5.1.

From these results it is important to retain that the best patterns include cases that consider section 2 of Fig. 5.5 (cases A12 and A21), that is, cases that include the terminal equipments in the calculations[2]. Such results are in line with those that were achieved for the local calling area scenario (subsection 5.2.2).

Notice also that both cases A21 and A12 consider that the reference signal is injected at the virtual 0 dBr point. Hence, the assigned values to SLR_A and RLR_B are relative to the difference between the reference signal level and the levels that were measured at side A and side B, respectively.

[2]The difference between A21 and A12 is that A12 considers the terminal equipments as being analog, whereas A21 consider them as being digital. This is reflected in the SLRa and RLRb values, that are $SLRa$ = +11 and $RLRb$ = −3 for the analog equipment and $SLRa$ = +7 and $RLRb$ = +3 for the digital equipment.

Alcatel System 12

The same procedure as the described in regard to the Siemens switch is applied to the circuit configuration that includes the Alcatel System 12 switch. Results are shown in Table 5.14.

For this switch it is not possible to establish an attenuation case for which all the MOS error values are less than a certain value. In fact, the cases for which the absolute MOS error value is $|MOS\ error| < 0.15$ range from A00 to A25 (see row 1 of the Table 5.14), giving that A21 and A12 are excluded, contrary to what happened with the Siemens switch, in which A12 and A21 played an important role on the MOS calculation accuracy. If A12 and A21 are added to this range of cases, the MOS Error increase such that $|MOS\ Error| \leq 0.17$. However, in this pattern, all the cases include *WEPL* values that

Table 5.14 – MOS errors obtained for each of the patterns, for the Alcatel switch

Input patterns		**|MOS Error|**	**Remarks**
#	**Combinations of cases**		
1	{A00, A01, A11, A12[3] , A13, A14, A15, A16, A17, A18, A19, A20, A21,A22, A23, A24, A25}; N00; D00; W00; {T01, T00.}	[0.0001; 0.17]	WEPL value out of the permitted range
2	A21; N00; D00; W01; T00	0.18	The best pattern combination. Predominance of default values.
3	{A00, A11, A15, A16, A19, A20, A22, A23, A24, A25}; N00, D00, W00, T00.	[0.18; 0.26]	WEPL value out of the permitted range
4	A12; N00; D00; W01;T00	0.29	Second best pattern combination. Predominance of default values.
5	A21; N00; D00; W02; T00	0.32	Third best pattern combination. Predominance of default values.
6	{A13, A14, A17, A18} N00; D00; W00; T00.	0.33	WEPL value out of the permitted range

[3]Cases A12 and A21 are, here, considered. This is the reason to have $|MOS\ error| < 0.17$ instead of $|MOS\ error| < 0.15$, as mentioned in the text.

are out of the permitted range, as recommended in the ITU-T Rec. G.107 [47]. This contingency results in an invalid pattern and suggests the need to increase the error tolerance so that a valid pattern can be found. After doing so, the best combination includes the A21 together with N00, D00, W01 and T00 relative to the noise, delay, Weighted Echo Path Loss and Talker Echo Loudness Rate values, respectively. It is the pattern of the row 2, and the absolute MOS error is $|MOS\ error|$ = 0.18. The error is 0.03 above the initial specification (Eq. 5.1), which is considered acceptable.

The next pattern to consider is shown in the row 4 of the Table 5.14. It comprises the cases A12, N00, D00, W01 and T00, to which corresponds $|MOS\ error|$ = 0.29. Notice that all the input values are default values, except the one used in A12.

Another valid pattern is presented in row 5. It is formed by the cases A21, N00, D00, W02 and T00, to which corresponds an error of $|MOS\ error|$ = 0.32. Also, in this pattern, the used values are predominantly default values.

Despite the fact that achieved results are not so conclusive as they are in the case of the Siemens switch, it is possible to conclude that the attenuation is the parameter that plays the main role and that the use of default values is acceptable to derive MOS values with relative accuracy.

Similar to the Siemens switch, the best case whose input parameters fall in the range permitted by the E-Model recommendation, include A12 and A21. These cases consider the reference signal as injected in the 0 dBr point and also the section 2 of the Fig. 5.5. However, in the case of a local calling area scenario (section 5.2), the corresponding path does not include the Loudness Rate values of the telephones, when this switch is considered.

Similar to the procedure used in the local calling area scenario, the error resulting from applying the Alcatel suitable cases to the Siemens switch as well as the error resulting from applying the Siemens suitable cases to the Alcatel switch is analysed. Table 5.15 summarises the obtained results.

As it can be seen in Table 5.15, the application of the suitable cases to respective switches leads to good results, since MOS errors are 0.055 and 0.18, respectively. When using foreign cases, there are the Alcatel cases that lead to the best results: an error of 0.45 against 0.52 if the suitable Siemens cases are applied to Alcatel. So, if for any reason only one pattern must be applied to both switches, the Alcatel suitable pattern leads to the best results.

Table 5.15 – $|MOS\ Error|$ that is possible to achieve from applying the same cases to both switches.

	Siemens suitable case	**Alcatel suitable case**
Siemens switch	0.055[4]	0.45
Alcatel switch	0.52	0.18[5]

A trade-off pattern for both switches

The approach based on the suitable/foreign patterns, whilst optimising results for one switch, dramatically degrades results relative to the other one. Aiming to find a unique trade-off solution that can be applied to both switches without degrading results so dramatically to one switch when optimising for the other, is now being identified. Table 5.16 shows the patterns that simultaneously lead to the lowest MOS errors for both switches. The third column shows the patterns and, at its left and right sides, relative MOS errors concerning Siemens and Alcatel switches are shown.

By analysing the table it is possible to observe that the better trade-off patterns are those of rows 38 and 36 (A12, A21, N00, D00 {W01, W02}, T00). In this case the resulting error is $|MOS\,Error| \lesssim 0.32$. Despite the fact that it is a relatively high value, it can be acceptable if it is taken as an error margin. For example, taking into account the ITU-T Rec. G.109, that establishes a minimum acceptable value of $R = 50$, this means a minimum acceptable $MOS = 2.6$ [142]. Since our error margin is 0.32, a system to be safely acceptable must have a MOS such that $MOS \geq 2.6 + 0.32 = 2.92$. The result given by this pattern corroborates the importance of considering the reference signal of 1 020 Hz, −10 dBm to be injected in the virtual 0 dBr point, combined with the Loudness

[4]The best case that uses measured values is supposed. (See row 3, Table 5.13).
[5]The best case that uses default values is supposed. (See row 2, Table 5.14).

Rate default values of the terminal equipments.

Aiming to find a trade-off pattern that uses values measured by ArQoS®, the patterns presented in the lines 58 and 59 (A22, N22, D11, W22, T22 and A23, N23, D11, W23 and T23) are found as the best ones. However, as it is possible to verify, the error may reach 0.8322. Thus, according to what has been previously discussed, considering this pattern must be restricted to the cases where MOS are $MOS > 2.6 + 0.822 = 3.422$.

Generically, for an error margin, $Error\ margin = |MOS\ error|$, the acceptable derived MOS value must be such that,

$$MOS > 2.6 + Error\ margin. \tag{5.12}$$

Table 5.16 – Patterns common to both switches with acceptable accuracy

#	Siemens MOS error	Patterns of inputs	Alcatel MOS error	Remarks
1	-0.0503;	A21; N21; D00; W21; T21	0.5417	
2	-0.0548;	A21; N00; D00; W02; T01	0.5344	
3	-0.0555;	A21; N21; D11; W21; T21	0.5238	
4	-0.0583;	A12; N00; D00; W01; T01	0.5032	
5	0.0617;	A12; N00; D11; W02; T01	0.6240	
6	-0.0666;	A21; N00; D11; W02; T01	0.5241	
7	0.0730;	A12; N00; D00; W02; T01	0.6335	
8	0.0793;	A12; N12; D11; W12; T12	0.6352	
9	0.0889;	A12; N12; D00; W12; T12	0.6585	
10	0.1392	A23; N00; D00; W01; T00	0.5378	
11	0.1562	A25; N00; D00; W01; T00	0.5369	
12	-0.1573	A01; N00; D00; W00; T01	0.0774	WEPL value out of the permitted range
13	0.1582	A22; N00; D00; W01; T00	0.5556	
14	-0.1621	A12; N00; D00; W02; T00	0.4303	
15	-0.1624	A13; N00; D00; W00; T01	-0.0781	WEPL value out of the permitted range
16	-0.1644	A14; N00; D00; W00; T01	-0.0759	WEPL value out of the permitted range
17	-0.1677	A17; N00; D00; W00; T01	-0.0765	WEPL value out of the permitted range
18	-0.1693	A18; N00; D00; W00; T01	-0.0743	WEPL value out of the permitted range
19	0.1781	A24; N00; D00; W01; T00	0.5547	
20	-0.1848	A21; N00; D00; W01; T01	0.4029	

Continued on next page

Table 5.16 – *Continued from previous page*

	Siemens MOS error	Patterns of inputs	Alcatel MOS error	Remarks
21	-0.1936	A00; N00; D00; W00; T01	0.0411	WEPL value out of the permitted range
22	0.2330	A11; N00; D00; W01; T00	0.7407	
23	-0.2418	A19; N00; D00; W00; T01	0.0399	WEPL value out of the permitted range
24	-0.2477	A15; N00; D00; W00; T01	0.0402	WEPL value out of the permitted range
25	-0.2487	A20; N00; D00; W00; T01	0.0342	WEPL value out of the permitted range
26	-0.2548	A16; N00; D00; W00; T01	-0.3577	
27	-0.2564	A16; N00; D00; W01; T00	0.6418	
28	-0.2639	A11; N00; D00; W00; T01	0.0656	WEPL value out of the permitted range
29	0.2721	A20; N00; D00; W01; T00	0.6409	
30	0.2747	A15; N00; D00; W01; T00	0.6582	
31	0.2846	A23; N00; D00; W02; T00	0.6819	
32	-0.2872	A24; N00; D00; W00; T01	0.0039	WEPL value out of the permitted range
33	0.2902	A19; N00; D00; W01; T00	0.6573	
34	-0.2942	A22; N00; D00; W00; T01	0.0043	WEPL value out of the permitted range
35	-0.2951	A25; N00; D00; W00; T01	-0.0027	WEPL value out of the permitted range
36	-0.2982	A21; N00; D00; W02; T00	0.3195	Predominance of default values
37	0.3016	A25; N00; D00; W02; T00	0.6809	
38	-0.3020	A12; N00; D00; W01; T00	0.2851	Predominance of default values
39	-0.3023	A23; N00; D00; W00; T01	-0.0024	WEPL value out of the permitted range
40	0.3036	A22; N00; D00; W02; T00	0.6994	
41	0.3205	A24; N00; D00; W02; T00	0.6984	
42	0.3486	A23; N00; D00; W01; T01	0.7277	
43	0.3638	A25; N00; D00; W01; T01	0.7269	
44	0.3655	A22; N00; D00; W01; T01	0.7431	
45	0.3780	A11; N00; D00; W02; T00	0.8775	
46	0.3806	A24; N00; D00; W01; T01	0.7423	
47	0.4011	A16; N00; D00; W02; T00	0.7831	
48	-0.4061	A01; N00; D00; W00; T00	-0.1714	WEPL value out of the permitted range
49	-0.4115	A13; N00; D00; W00; T00	-0.3332	WEPL value out of the permitted range
50	-0.4136	A14; N00; D00; W00; T00	-0.3309	WEPL value out of the permitted range
51	0.4167	A20; N00; D00; W02; T00	0.7823	
52	-0.4170	A17; N00; D00; W00; T00	-0.3315	WEPL value out of the permitted range
53	0.4187	A18; N00; D00; W00; T00	0.3292	WEPL value out of the

Continued on next page

Table 5.16 – *Continued from previous page*

	Siemens MOS error	**Patterns of inputs**	**Alcatel MOS error**	**Remarks**
				permitted range
54	0.4192	A15; N00; D00; W02; T00	0.7989	
55	0.4318	A11; N00; D00; W01; T01	0.8985	
56	0.4345	A19; N00; D00; W02; T00	0.7981	
57	-0.4349	A21; N00; D00; W01; T00	0.1754	
...				
58	0.4694	A22; N22; D11; W22; T22	0.8322	The best pattern that uses measured values
59	0.4694	A23; N23; D11; W23; T23	0.8322	The best pattern that uses measured values

The results and discussion presented so far, in subsections 5.2.2 and 5.3.3, presented solutions in which input patterns include the use of both default values and measured values. Each time a solution using default values was presented, a solution using measured values was also presented. The achievement of better results when using default input parameter values can be seen as consequence of carrying out tests under almost ideal conditions, where noise and delay were virtually null, not influencing results. However, whenever possible, measured values should be used, since they represent the real conditions and so, more reliable results can be achieved. In this case, a more restrictive condition for the derived MOS must be observed to ensure the required accuracy. This can be achieved by using the expression 5.12.

The investigations carried out permitted to implement the voice quality evaluation model, as described in Appendix A, in the form of algorithms: Algorithm 1, Algorithm 2 and Algorithm 3.

5.4 The VoIP module

The modules presented so far cannot be applied to the packet switching technology neither to VoIP communications since they do not take into account VoIP inherent factors such as packet loss. To evaluate VoIP systems, the experimental study presented in this section has been carried out so that a VoIP evaluation module can be derived.

To derive the VoIP module, three main ITU-T recommendations for voice quality evaluation are taken into account:

- The ITU-T Rec. G.107, that describes the "E-Model", the chosen basis for deriving the non-reference model as discussed in section 5.1 [47].
- The ITU-T Rec. P.564, ("Conformance testing for voice over IP transmission quality assessment models"), that specifies the minimum criteria for objective speech quality assessment models that predict the impact of observed IP network impairments on the one-way listening quality experienced by the end-user in VoIP applications (3.1-kHz) [160].
- The ITU-T Rec. P.862, that describes the PESQ objective method for speech quality assessment of narrow-band telephone networks and speech codecs. As referred to in section 5.1, it constitutes the reference for validation of the derived model, as required by the ITU-T Rec. P.564.

In this trial, the impairments caused by both low bit-rate codecs and voice packet losses of random distribution are under study. Thus, in the E-Model expression, $R = R_0 - I_s - I_d - I_{e-eff} + A$ (see Eq. 3.5), special attention was paid to the term I_{e-eff}, which represents these type of impairments, as referred to in section 3.5. The validation of the derived model is done according to the conformance testing procedures described in the Rec. ITU-T P.564 [160]. The use of PESQ as reference is additionally justified by the fact that both the E-Model and and PESQ are sensitive to distortions caused by codecs and packet loss. It is necessary that the impairment factors taken into account are common to both models; otherwise PESQ could not be the best reference model to calibrate the derived model.

5.4.1 Adjusted methodology

The test scenario presented in Fig. 5.6 was used in this study, where the main signal path includes coding and packetization and random packet-loss in an IP Network as well as decoding, from which the degraded signal is obtained. Thereafter, on the one hand, both reference and degraded signals are given as inputs to the PESQ algorithm, whilst

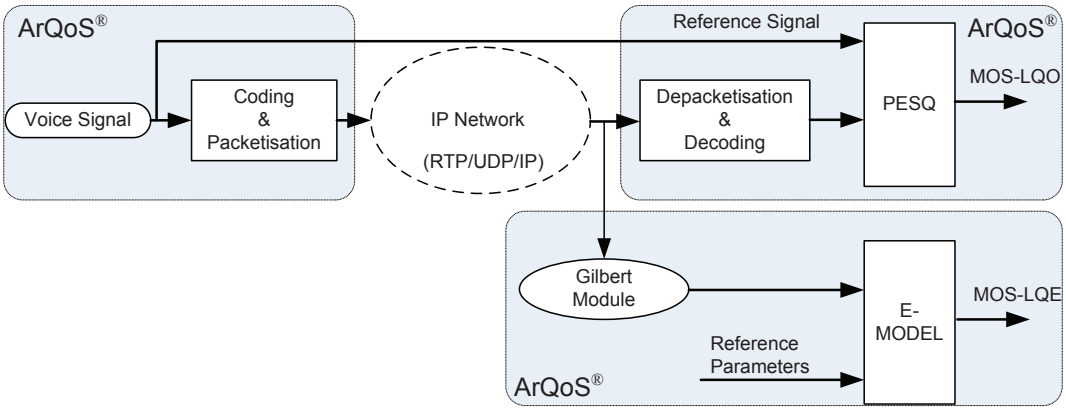


Figure 5.6 – Experimental setup for validation and calibration of the E-Model.

the output is the reference MOS used to calibrate the derived model. The used MOS is the MOS_{LQO}[6] as recommended by the ITU-T Rec. P862.3 [161]. On the other hand, the degraded voice stream was collected and applied to a Gilbert modeling module whose output gives the probabilities necessary to calculate the Ppl and $BurstR$ values for I_{e-eff} (See the meaning of Ppl and $BurstR$ parameters in subsection 3.5.1, pages 61 and 62). Concerning the distortion caused by codecs, respective impairment factor values are taken from the table 2a/G.108 of ITU-T Rec. G.108 [141].

The experimental tests have been carried out in two stages:

- The first stage aims at achieving an accurate voice quality model. The voice samples defined in Rec. ITU-T P.501 were used in the tests [162]. Two male and two female speaker sentences were used, comprising English and Spanish languages. They were downsampled to 8 kHz (16 bits) as required by PESQ and recommended by ITU-T Rec. P.862.3 [161]. Table 5.17 shows the samples used in this calibration stage.
- The second stage is aimed to validate the results obtained in the first stage by using a new set of sentences and carrying out new experiments. The test scenario and the test conditions are the same as in the calibration tests described above. Table 5.18 shows the test sentences used in this validation stage.

[6]Listen Quality Objective MOS

Table 5.17 – Sentences used in the first stage of the trial.

Test sentences	Gender	Language
These days a chicken leg is a rare dish. The hogs were fed with chopped corn and garbage.	Female 1	English
The juice of lemons makes fine punch. Four hours of steady work faced us.	Male 1	English
No arroje basura a la calle. Ellos quieren dos manzanas rojas.	Female 1	Spanish
P - siéntate en la cama. El libro trata sobre trampas.	Male 1	Spanish

Table 5.18 – Used sentences on the validation stage

Test sentences	Gender	Language
Rice is often served in round bowls. A large size in stockings is hard to sell.	Female 2	English
The birch canoe slid on smooth planks. Glue the sheet to the dark blue background.	Male 2	English
No cocinaban tan bien. Mi afeitadora afeita al ras.	Female 2	Spanish
El trapeador se puso amarillo. El fuego consumió el papel.	Male 2	Spanish

The codecs used in the trials for evaluation and calibration were the commonly used VoIP codecs G.711, G.729 8 kbps and G.723.1 6.3 kbps [23]. Six average packet loss ratios were selected to take the relevant results: 0%, 2.5%, 5%, 10%, 15% and 20% [23]. The MOS_{LQO} values obtained from PESQ, as well as those obtained from the modified E-Model were collected for each packet loss rate, codec and sentence. This resulted in a total of 24 tests for each codec and 24 different MOS scores for each evaluation method, i.e, 24 MOS scores for the modified E-Model and 24 MOS scores for PESQ for each codec. Then for each codec, regression analysis was used to calibrate the derived voice quality model. Based on these two sets of scores (from PESQ and from modified E-Model), the coefficients of a polynomial $p(x)$ that fits *p(E-Model MOS)* to MOS_{LQO} were derived.

5.4.2 Results and discussion

Fig. 5.7 shows the results obtained from regression analysis done in the first stage, that models the relationship between MOS_{LQO} and the derived model MOS scores for the G.711 codec. The horizontal axis contains the scores obtained from the derived model, while the vertical axis represents the scores obtained from PESQ (MOS_{LQO}). For each point in the graph, the difference between the scores is the error between the modified E-Model and the reference PESQ. For instance, the second point from the left corresponds to E-Model MOS=1.5 and MOS_{LQO}=1.8, which means a MOS error of 0.3. In this case, the E-Model underestimates the MOS score in comparison with PESQ. In the graph, the points over the straight line correspond to no error cases in which both models produce the same result. In general, this figure shows that the derived model overestimates MOS relatively to PESQ. Therefore, a function to approximate the E-Model output to that of PESQ was derived. The figure shows also the trend line that minimises the RMSE between both MOS scores, which is the polynomial line that best approximates the E-Model to PESQ. Such line corresponds to the coefficients of a polynomial of degree 4, which gives the best approximation to PESQ. The resulting polynomial is given by

$$\begin{aligned} MOS_{LQO} \;=\; & -0.0058 MOS_{LQE}^4 + 0.1252 MOS_{LQE}^3 - \\ & -0.6467 MOS_{LQE}^2 + 1.9197 MOS_{LQE} - \\ & -0.291, \end{aligned} \tag{5.13}$$

which is the calibrating function of the E-Model MOS in order to get the corresponding MOS_{LQO} scores from the E-Model MOS scores.

Fig. 5.8 shows the MOS scores obtained for the G.729 codec under the same test conditions as described in the previous case. The figure shows that in this case, the derived model overestimates the MOS, when compared with MOS_{LQO} from PESQ. Fig. 5.8 also shows the trend line that best approximates the E-Model scores to MOS_{LQO} from PESQ algorithm. For this codec, the polynomial function to approximate the E-Model results to those of

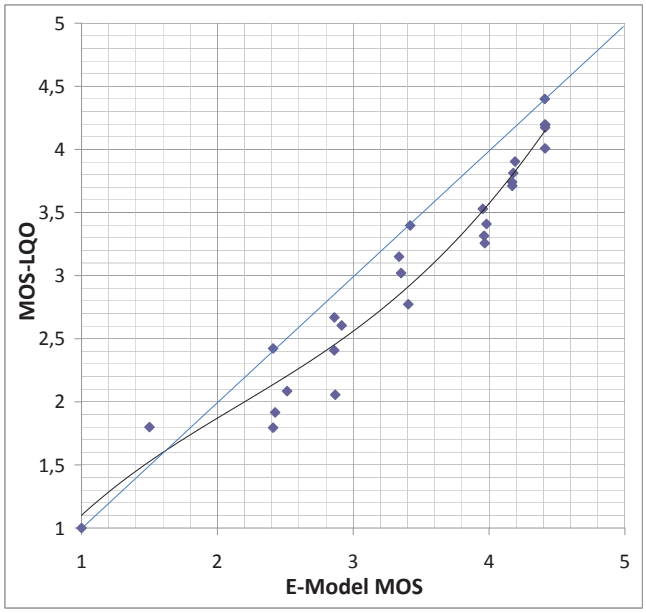


Figure 5.7 – Regression modeling of E-Model MOS scores as MOSLQO for G.711

PESQ MOS_{LQO} is given by

$$\begin{aligned} MOS_{LQO} &= -0.0554MOS_{LQE}^5 - 0.7496MOS_{LQE}^4 + \\ &\quad +3.9507MOS_{LQE}^3 - 9.874MOS_{LQE}^2 + \\ &\quad +11.939MOS_{LQE} - 3.8293. \end{aligned} \tag{5.14}$$

Finally, Fig. 5.9 shows the results for G.723.1 codec. In this case, the derived model underestimates MOS. The figure also shows the polynomial trend line that best approximates the E-Model scores to MOS_{LQO} from PESQ algorithm.

From these results, the function that best approximates MOS from E-Model to PESQ is given by:

$$\begin{aligned} MOS_{LQO} &= 0.0018MOS_{LQE}^4 + 0.0248MOS_{LQE}^3 - \\ &\quad -0.4262MOS_{LQE}^2 + 2.1953MOS_{LQE} - \\ &\quad -0.2914. \end{aligned} \tag{5.15}$$

At this first stage a model is derived by have deriving these three functions; one for each codec. Next stage is concerned with validation of the derived functions. In this stage,

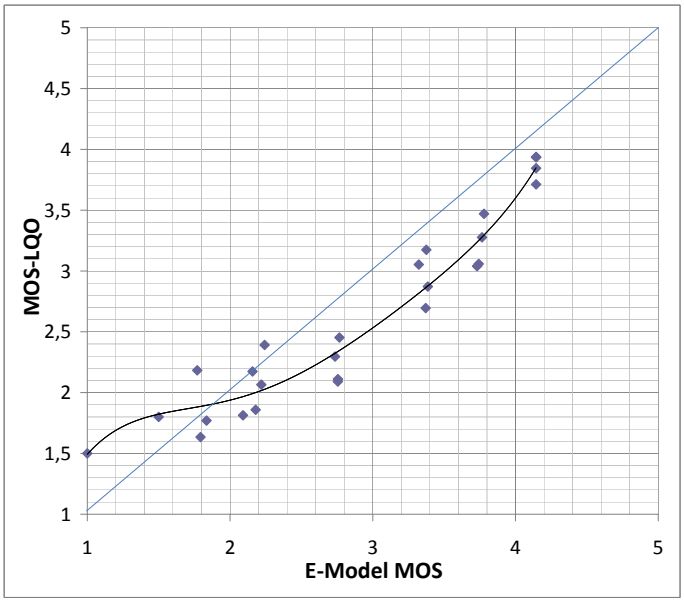


Figure 5.8 – Regression modeling of E-Model MOS scores as MOSLQO for G.729

the sentences of Table 5.18 were used in the ArQoS® test system to obtain the respective MOS_{LQO} and the derived model MOS scores, now calibrated by using Equations 5.13, 5.14 and 5.15.

As ITU-T Rec. P.564 defines and requires, the correlation factor, error and false positive/negative analysis between MOS_{LQO} scores and MOS obtained from the derived model are determined. Table 5.19, Table 5.20 and Table 5.21 show the results obtained from the tests and the conformance accuracy requirements defined in ITU-T P.564. These tables show the correlation factor, percentage of errors and false negative/false positive measures, respectively.

As it can be seen in Table 5.19 and Table 5.21, the obtained results match both the correlation and false negative/false positive requirements for the Class 1 of accuracy specified in ITU-T Rec P.564, except when using the codec G.723.1, since respective correlation value (0.887) does not match the Class C1 requirement (>0.900). Nevertheless, it matches the Class C2 requirement (>0.850). One could say that our derived model falls into the Class C1 given that only G.711 and G.729 codecs are used. However, according to the results shown in Table 5.20, the percentage of errors falls within boundaries 1, 7 and 8

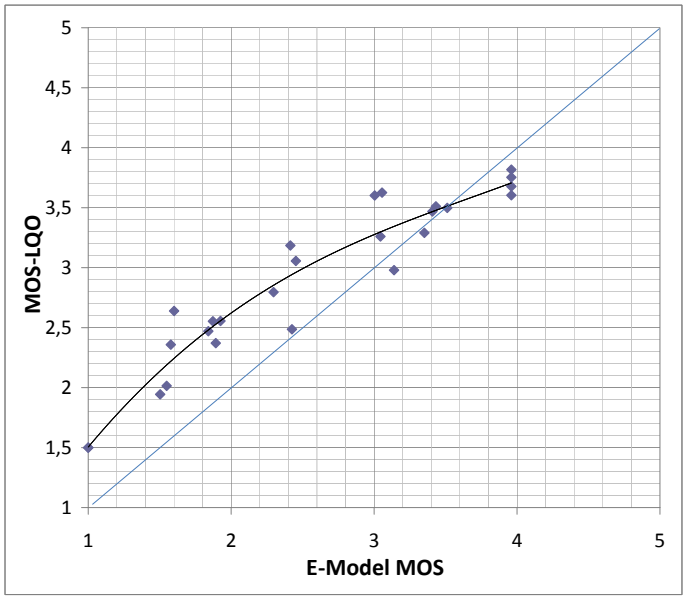


Figure 5.9 – Regression modeling of E-Model MOS scores as MOSLQO for G.723.1

(even for codecs G.711 and G.729), which makes the derived model to be included into Class 2.

Given that ITU-T Rec. P.564 requirements are met, the derived voice quality evaluation model module was integrated and tested in the passive probes of ArQoS®.

Table 5.19 – Results for the correlation factor

Measures	Results			Requirements (P.564)	
	G.711	G.729	G.723.1	Class C1	Class C2
Correlation	0.956	0.964	0.887	> 0.900	> 0.850

Fig 5.10 shows the context in which this model operates. As it is depicted, the ArQoS® passive probes are deployed in the Portugal Telecom VoIP Network core. All Real-time Transport Protocol (RTP) streams are transmitted through the core, either in calls between VoIP and circuit-switch endpoints, or between just two VoIP clients. The derived model is applied in every call from which two MOS calculations are performed, one for each way. On this application scenario, the network problems that affect the RTP stream after its passage through the core are not really detected by the probes. It is the reverse RTP stream that follows the same path, and is affected to some extent, that is analysed

Table 5.20 – Results for the percentage of errors

Errors within standard bounds	Results			Requirements (P.564)	
	G.711	G.729	G.723.1	Class C1	Class C2
Quality band B=1 ($MOS_{LQO} \geq 2.8$)					
Boundary 1 (%)	81	90	67	≥ 95.0	≥ 75.0
Boundary 2 (%)	100	100	100	≥ 97.9	
Boundary 3 (%)	100	100	100		≥ 95.0
Boundary 4 (%)	100	100	100	≥ 99.0	
Boundary 5 (%)	100	100	100		≥ 97.9
Boundary 6 (%)	100	100	100		≥ 99.9
Quality band B=2 ($MOS_{LQO} < 2.8$)					
Boundary 7 (%)	75	86	78	≥ 90.0	
Boundary 8 (%)	88	100	89		≥ 90.0
Boundary 9 (%)	100	100	100	≥ 95.0	
Boundary 10 (%)	100	100	100		≥ 95.0
Boundary 11 (%)	100	100	100	≥ 99.0	
Boundary 12 (%)	100	100	100		≥ 99.0

Table 5.21 – Results for false negatives and false positives

Measures	Results			Requirements (P.564)	
	G.711	G.729	G.723.1	Class C1	Class C2
False negatives (%)	0	0	0	< 5	< 5
False positives (%)	0	0	0	< 3	< 3

by the probes. The calculated MOS values are also processed and shown in the ArQoS® statistics reporting tool, giving the users a good overview of the network voice quality.

PSTN network is also referred to in the figure. It is connected to the VoIP network via a Media Gateway that acts as a bridge between the different used technologies. It is in the PSTN network where the modules derived in sections 5.2 and 5.3 apply.

The investigations carried out so far permitted implement the voice quality evaluation model, as described in Appendix A in the form of an algorithm (Algorithm 4).

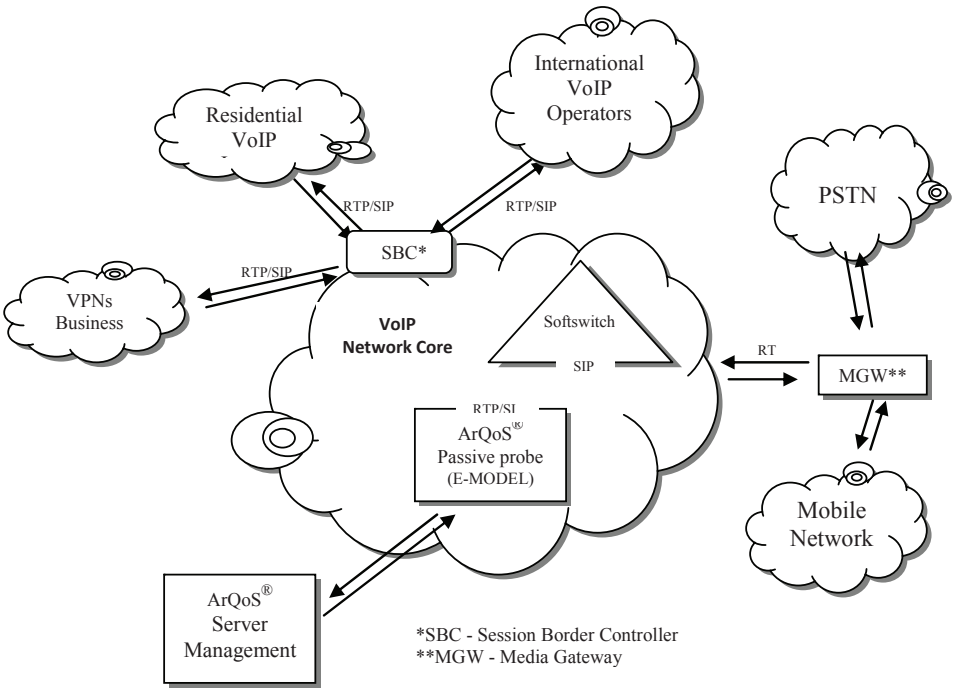


Figure 5.10 – Portugal Telecom VoIP network (Courtesy of Portugal Telecom Inovação)

5.5 Conclusions

In this chapter a practical model to evaluate the voice quality in the telephony context has been described. It encompasses the circuit switched local calling area, the far-end calling area and the VoIP scenarios.

Section 5.1 describes the methodology that was formulated to be used on deriving the aimed model. It describes how the E-model was chosen to base the derived model and the reasons that support that choice as well as the choice of the PESQ to be the reference against with achieved results are confronted. A call quality monitoring system (ArQoS®) has also been presented as means of giving input parameters to the derived model. A preliminary study making use of this system and the given parameters has been described. Preliminary results dictated the need to further study the nature of the available parameters in order to identify which of them best fit the needs of the model under study. This procedure was then added to the methodology to be applied in both the local and far-end calling area scenarios as described in sections 5.2 and 5.3, respectively.

By applying the formulated methodology, input parameters that minimise the MOS error between the reference and the derived model have been identified in section 5.2. Achieved results shown that MOS accuracy mainly depend from loudness rates and Weighted Echo Path Loss values for the test conditions. The derived model fits both Siemens EWSD and Alcatel System 12 models.

The study presented in section 5.3 is similar to this of section 5.2, except that it applies, now, to the scenario in which two switches are connected via the SS7 signaling system, so that a far-end scenario is simulated.

Section 5.4 describes the derived module concerning the evaluation of the voice quality in a packet switching context, as is the VoIP. It relies essentially on considering the packet loss and codec distortion as the main impairment factors. A Gilbert modeling module is used to derive the packet loss probability and burst ratio needed inputs from the voice stream. Concerning codec distortion, values tabulated in ITU-T Rec. G.108 are used. The derived model is validated according to ITU-T Rec. P. 564 requirements. It complies with the specified class C2 of accuracy. Hence, it was tested and integrated for production in Portugal Telecom Comunicações.

6 MOS enhancement by packet classification-and-prioritisation

This chapter presents a research study on voice packet classification and prioritisation according to the relevance of their payload to the overall voice quality. An algorithm to classify voice packets according to their importance to perceptual quality is proposed with the aim of establishing different priorities and then drop first those of lower importance in congested networks to implement different priorities.

After an introductory review, section 6.1 describes the proposed algorithm that optimally classifies voice packets according to a distortion minimisation criterion. Then, section 6.2 briefly describes two random packet loss modeling algorithms used in the experiments carried out in the simulation study presented in section 6.3. In this section, the results are also presented and discussed in the context of a generic application scenario where voice packets can be transmitted with different priorities. Section 6.4 proposes a combined technique that uses the classification and the Papoulis-Gerchberg algorithms in a tandem arrangement to enhance reconstruction performance.

6.1 Voice packet prioritisation

Prioritisation of voice packets is useful to cope with the best-effort nature of Internet, where UDP is used to transport voice packets with no delivery guarantee within the necessary time limits. Thus, depending on the network conditions, random packet losses

are likely to occur in a random manner, without taking into account any possible difference between packets carrying data with different levels of importance.

On the one hand, one may argue that packet losses are not totally harmful, since small losses are imperceptible to the human ear. Nevertheless, beyond a certain loss rate, they tend to affect the voice quality and the conversation intelligibility [50]. On the other hand, not all missing packets have the same impact on the voice quality degradation, since this depends on the packet content, which in turn is reflected by the position of its payload in the speech signal. Fig. 6.1 shows how the MOS can vary as a function of the packet loss location. As it can be seen, not all packet losses have the same impact on MOS, specifically on decreasing it. For example, losing the 40^{th} packet decreases the MOS from $\gtrsim 4.5$ to 3.5 whereas losing the 80^{th} packet decreases the MOS to $\lesssim 4$. A study about the different packet importance in voice quality is presented in [95].

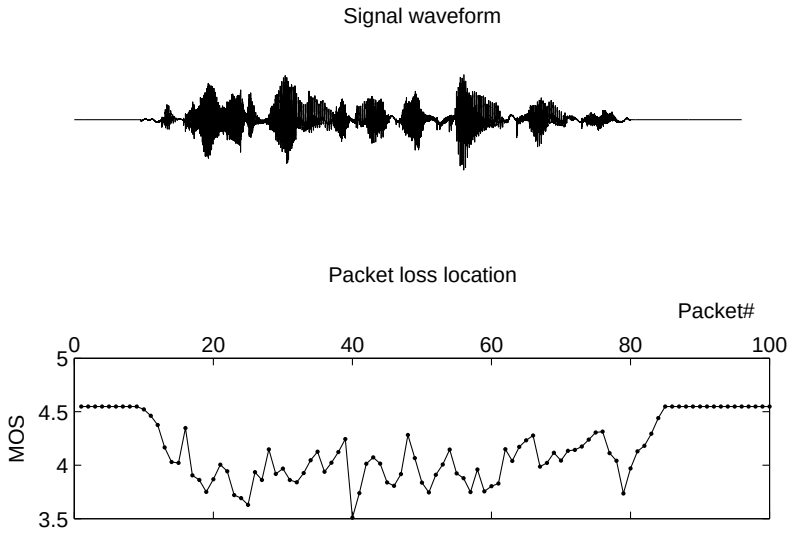


Figure 6.1 – Impact of the packet loss location on the voice quality degradation [95]

The effects of packet loss can be minimised by using recovering techniques such as signal reconstruction referred to in previous chapters. However, since not all packets contribute with the same importance to the overall voice quality and losses occur randomly, packets

with quite different importance levels are indistinctly lost. The effect of such random losses can be overcome by judiciously classifying the packets at the source such that more importance is given to those that contribute more to the perceived quality. Specifically, these different importance levels can be used for differentiating the most important packets in priority-enabled networks. The underlying idea is to classify the voice packets as either less important and more important to intelligibility and give more priority to the most important packets in order to reduce the loss probability of high importance packets in congested networks. In a communication channel where data flows permit to assign high and low priorities, the low priority packets are always the first ones to be discarded (in case of network congestion, for example) or are those that are less protected (with error correction codes, for example). As a result, a better voice quality is expected to be achieved in comparison with the case where no packet differentiation is used. Moreover, with such a classification scheme, eventual signal reconstruction can be more effective because it relies, *a priori*, on the most important packets.

In this context, it is appropriated to classify these packets as high priority and low priority packets. Signal reconstruction can be done by simply substituting the missing packets by the previous last known one, that is, by performing a zero-order hold interpolation or using a more sophisticated method, such as those presented in chapter 4.

To carry out packet prioritisation it is necessary to classify each predetermined voice chunk within a longer sample sequence according to its importance. Due to the fact that the percentage of packets to prioritise may vary and, for a given percentage, several combinations of high/low priority packets can be found, there is a huge number of solutions for the prioritisation problem, where some of them are more efficient than others. In this work the most efficient solution is called the optimal solution.

The classification algorithm proposed in the next sections aims to find the optimal solution. It is inspired in video summarisation, where a short subset of video frames is selected as the most representative of a long sequence [163]. In the case of voice packet optimal classification, the aim is to find a subset of packets that minimise the distortion of the reconstructed signal when the least important packets are lost.

It is worth to mention that losses and classification are independent processes. The classification process is deterministic and classifies each chunk with the same importance – the importance it has on the voice intelligibility of that sentence. The loss process is stochastic and only depends on the network conditions, which are continuously varying. This means that not all low priority packets are necessarily lost and that high priority ones are likely to be lost, too. But, on average, high priority are less likely to be lost.

Problem formulation

To formulate the optimal packet classification problem, several concepts and terminology are defined as follows:

- **Utterance**: consists in a speech sample sequence of one or more vocal sounds preceded and followed by silence.
- **Segment**: chunk of voice signal samples belonging to a utterance. One utterance is comprised of several segments.
- **Packet**: network layer data unit encapsulating one segment as payload.
- **Priority rate**: ratio between the number of high priority packets and the total number of packets comprising a utterance.

Denoting by n the total number of packets in a utterance, U, and by m the number of high priority packets in subset, M, we define the priority rate, $R(M) \in]0, 1]$[1], as follows:

$$R(M) = m/n. \tag{6.1}$$

The problem of optimal packet classification into low and high priority is formulated as follows: *Given a utterance of n segments, corresponding to n packets, and a maximum priority rate $R(M)$, the problem is to find the best $m \leq n$ segments that minimise some distortion function, $D(m,n)$, between the original utterance and the one reconstructed from the reduced subset, M, of m segments.*

The segments in subset M are those that minimise the reconstruction distortion, hence they correspond to the payload of the high priority packets. These are the ones that carry the most important information for the best reconstruction, when either some of all of

[1] As it shall be explained later, the first segment, p_0, always belongs to the m-subset.

the others (*i.e.*, low priority) are lost in the network. Thus, the $n-m$ low priority packets are those that can preferentially be discarded, whenever necessary.

Each missing packet produces a sequence of sample erasures in the signal. Then, the original utterance of length n must be reconstructed from the m known segments, given that the erasure positions are known. Therefore, the reconstructed voice signal is always a distorted version of the original signal.

To mathematically define the problem, the following further definitions are necessary. Let the original voice signal (*i.e.*, the original utterance) be a temporal sequence of n segments defined as

$$U = \{p_0, p_1, \ldots, p_{n-1}\} = \{p_i\}\,,\ i \in \{0, 1, \ldots, n-1\}\,, \tag{6.2}$$

where p_i denotes the segment at position i.
Let the subset $M \subset U$ be defined as

$$M = \{p_{l_0}, p_{l_1}, \ldots, p_{l_{m-1}}\} = \left\{p_{l_q}\right\}\,,\ q \in \{0, 1, \ldots, m-1\}\,, \tag{6.3}$$

where l_q denotes the q^{th} segment that was selected from the original sequence, U, into subset M. We call this subset as the m-subset since its cardinality is m. The determination of the set of indices,

$$L = \{l_0, l_1, ...l_{m-1}\} = \{l_q\}\,, \tag{6.4}$$

completely defines this subset.

For the purpose of this work, original positions i not belonging to $\{l_q\}$ represent positions where segments are missing, which correspond to low priority packets. If all the low priority packets are lost, then the corrupted signal (*i.e.*, observed signal) is formed by the corresponding erasures plus the segments $\left\{p_{l_q}\right\}$.

As referred to above, we assume that zero-order hold interpolation is used at the receiver to reconstruct the signal such that each missing segment is substituted by the last known one. This interpolation was chosen since it is of trivial computation while still serving the purposes at this stage of the research work. However, any other reconstruction methods can be used, namely those presented in chapter 4.

If all m high priority packets are transmitted and all $n-m$ low priority packets are lost, then erasures are concealed through substitution by the last high priority (known) packet. However, this is an extreme case, since in a practical situation the network behaviour is not so straightforward, because of the random nature of packet losses. Nevertheless, this is a useful case to consider for the explanation of the problem and its solution.

Example

The following example provides a more detailed description of the above concepts and the problem definition. Without loss of generality, let us consider an utterance where $n = 5$ and a priority rate $R(M) = 3/5$, so $m = 3$. The corresponding original utterance is $U = \{p_0, p_1, p_2, p_3, p_4\}$.

Assuming that p_1 and p_4 are the less importance segments to be packetised as low priority, the sequence $C = \{p_0, p_2, p_3\}$ contains the complementary subset M of more important segments, whose indices are

$$\{l_q\} = \{0, 2, 3\}\,. \tag{6.5}$$

Considering that p_1 and p_4 are lost (*i. e.*, all low priority segments), by using zero-order hold interpolation, the reconstructed utterance, U', is given by,

$$\begin{aligned} U' &= \{p_0, p_0, p_2, p_3, p_3\} = \{p'_0, p'_1, p'_2, p'_3, p'_4\} \\ &= \left\{p'_j\right\},\ j \in \{0, 1, \cdots, 4\}. \end{aligned} \tag{6.6}$$

From Eqs. 6.2 and 6.6 it can be seen that p'_j is equal to segment p_i such that i is the greatest l_q lower than or equal to j. This means that any segment in the reconstructed utterance is equal to another one in the original utterance. The segment index j (*i.e.*, position) in the reconstructed utterance U' is always greater or equal than its original position in utterance U. That is,

$$p'_j = p_i : i = \max\{l_q\} \ \wedge\ l_q \leq j. \tag{6.7}$$

The distortion between two segments, u and v, is denoted by $d(u, v)$ and the distortion of the reconstructed signal $D(U')$, is defined as

$$D(U') = \frac{1}{n}\sum_{j=0}^{n-1} d(p_j, p'_j), \tag{6.8}$$

where p_j represents the j^{th} segment of the original sequence and p'_j represents the j^{th} segment of the reconstructed sequence.

Several different metrics may be used to measure the distortion. In this work the RMSE is used. It has the form:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(u[i] - v[i])^2}{N}}, \tag{6.9}$$

where $u[i]$ and $v[i]$ represent the i^{th} sample in segments u and v, respectively, and N represents the total number of samples in each segment.

For the specific example under analysis, $D(U') = \frac{1}{5}\Big[d(p_0, p'_0) + d(p_1, p'_1) + d(p_2, p'_2) + d(p_3, p'_3) + d(p_4, p'_4)\Big]$. Thus, the distortion of the utterance reconstructed from the reduced set defined by Eq. 6.5, is given by $D(U') = \frac{1}{5}\Big[0 + d(p_0, p_1) + 0 + 0 + d(p_3, p_4)\Big]$, since $d(p_j, p_j) = 0$.

Since the priority rate can be defined as an external parameter (*e.g*, user-defined, network driven, *etc.*), the following considerations should be taken into account. On the one hand, high priority rates tend to yield better reconstructed voice signals than low priority rates, in the case where the network conditions permit to forward all high priority packets. However, if this is not possible, there will be losses in high priority packets, which increases the reconstruction distortion because perceptually important packets are lost. On the other hand, if the priority rate is low, it is more likely that all high priority packets are received but voice quality may be worse than in the case of higher priority rate because utterance will mainly be formed by less perceptually important packets. Therefore, for a given maximum packet loss probability, there is a trade-off between the voice quality obtained from the reconstructed signal and the number of packets classified as high priority. However, the determination of the convenient priority rate value is not enough to solve the whole problem. Since $m < n$, there are several combinations of m high priority packets that are possible to be identified in the whole set of n packets representing the whole utterance. As defined before, this work aims to determine the best combination of m representative packets that ensures the minimum distortion when all the remaining ones are lost.

6.1.1 Sub optimal solution - greedy algorithm

If an exhaustive search is used to find all subsets of m segments contained in the original utterance of size n, it is necessary to deal with huge number of possible combinations, which is given by

$$\binom{n-1}{m-1} = \frac{(n-1)!}{(m-1)!\,(n-m)!} \quad [2]. \tag{6.10}$$

For high values of n and m, distortion computation of so many segments may be prohibitive. For example, to find all possible m-subsets in a sequence of 50 segments using $R(M) = 1/2$, results in $\binom{49}{24} \approx 6 \times 10^{13}$ different subsets that must be taken into account, which results in $\approx 6 \times 10^{13} \times 50 \approx 3 \times 10^{15}$ distortions to be calculated! Furthermore, additional computation is needed to determine which combination corresponds to the minimum distortion. Thus, exhaustive search may not be a practical solution and a different approach must be used in order to find a useful implementation at a reasonable computational cost.

In order to understand the structure of an alternative solution, we may consider an heuristic greedy algorithm [164, 165]. Since the aim is to minimise the distortion between the original and the reconstructed signal from a subset and, in general, greedy algorithms need a candidate set from which a solution is going to be iteratively created, the first iteration must have at least one segment in the target subset. Having only one segment in the initial subset corresponds to almost the maximum overall distortion subset[3]. Let the algorithm start with segment p_0 into the subset, that is, $p'_0 = p_0$. Thereafter, it computes all packet-by-packet remaining distortions, $\left\{d(p_j, p'_j)\right\}$, and selects into the subset the specific segment p^*_j, that corresponds to the maximum distortion. That is, $p'_j = p^*_j = \arg\max_{p_j \in U} \left\{d(p_j, p'_j)\right\}$. Since p^*_j corresponds to the maximum distortion, by including it in the subset, yields the maximum distortion reduction. Thus, this selection corresponds to "maximise the decrease" of distortion at each iteration.

In each of the further iterations, a similar distortion-based selection is performed to find the remaining segments to include in the subset. Therefore, at any iteration, t, the segment that maximises the distortion reduction is added to the $(t-1)$-subset. This process is

[2] $(n-1)$ and $(m-1)$ result from the fact that segment p_0 always belong to the m subset.
[3] The maximum overall distortion occurs when the subset is empty...

repeated until m segments are selected into the final subset, which becomes the m-subset. Note that such selection process can only find a local minimum because at each iteration the past selections cannot be modified, as it would be necessary in some cases to achieve the global minimum. Fig. 6.2 shows the processing structure of the greedy algorithm.

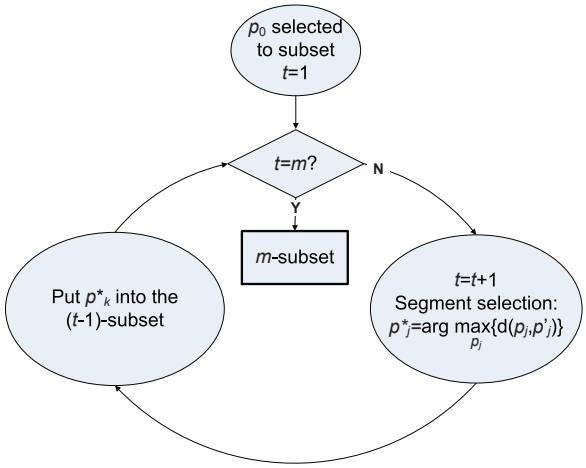


Figure 6.2 – Iterative operation of a greedy algorithm

Such greedy algorithm is better than the exhaustive search inherent to Eq. 6.10 because much less computations are required. However, as mentioned above, the resulting solution is suboptimal because each segment is selected without taking into account the final subset as a whole. A greedy algorithm always makes the choice that is best at the moment but not necessarily the best one for whole subset. That is, it makes a locally optimal choice in the hope that this choice will lead to a globally optimal solution [165]. Therefore, future decisions tightly depend on the past ones and thus, the global solution is suboptimal since a certain choice, in a certain past iteration, would be different if, at that iteration, all the different possible choices in the future could have been known. This means that, after the whole solution is found, one may conclude that some past choices may had not been the best ones.

Although greedy algorithms do not always yield optimal solutions, the previous example is useful to understand the nature and structure of our problem and provides a basis to understand the need of an alternative and the alternative.

A better strategy to solve the problem would not need to compute all combinations given by Eq. 6.10 and decisions taken in each iteration should not depend uniquely on the previous one(s). In other words, current decisions might be taken not only with knowledge about the past, but also with knowledge about the future (*i.e.*, segments ahead of the current ones). For this kind of optimisation problems, that can be separated by past, present and future and where, given the present, the future is independent of the past, Dynamic Programming (DP) is considered a powerful tool [166, 167]. This is the strategy presented in subsection 6.1.2 and the basis of the algorithm to optimally classify the segments of utterances.

6.1.2 The Dynamic Programming approach

With the concepts previously defined and the packet classification problem formulated as a rate-distortion optimisation problem [163, 168, 169], the optimal m-subset, M^*, is given by the solution of the following problem:

$$M^* = \arg\min_{M} D(U') \ : \ R(M) \leq R_{\max}. \tag{6.11}$$

This optimal subset M^* given by Eq. 6.11 can be obtained by searching over all the possible m-subsets, $M = \{p_{l_0}, p_{l_1}, ..., p_{l_{m-1}}\}$, containing m packets. By using DP, iterations are replaced by stages, t, in the process of looking for the global optimal solution. To each stage corresponds an intermediate subset, which cardinality is growing with the number of stages, similar to what happens in each of the greedy iterations. These intermediate subsets of size t are defined as t-subsets. In the case of DP, several possible segment candidates for the optimal m-subset are first identified at each stage, t. Then the best candidate is only selected in a second phase, taking into account not only the distortion reduction achieved by inserting the candidate itself into the subset, but also the differential distortion with candidates from the higher order adjacent stage.

By firstly identifying several segments as the possible optimal candidates without definitively selecting any specific ones into the optimal m-subset, it is possible to gather all of them until the last stage is reached. This leaves the final decision open while the process evolves through the various stages, allowing to come back later to determine which of the

candidates should be definitely selected from each stage. Therefore, the optimal segment to be chosen at each stage is dependent on the candidates identified in future stages. This is done in the second phase of the DP algorithm. This is quite different from the greedy algorithm, where one segment is definitely selected at each iteration, without knowing whether it belongs to the global optimum or simply to a local minimum.

To find the candidates at each stage, t, the algorithm identifies several segments $\{p_k\}$ from $U = \{p_i\}$. Since this is a sequential process, at each stage, t, the range of possible k must guarantee that previously selected segments are never chosen again and there are enough segments left to fulfill all subsequent stages, also with different segments from the previous ones.

Then, each segment p_k is inserted at the end of multiple $(t-1)$-subsets, which results in several different t-subsets associated with each p_k. The minimum distortion over all these t-subsets, defines the distortion state of each p_k. Therefore, at each stage, t, one single distortion state is found for each segment p_k, denoted by D_t^k.

The method described above is comparable to identify all the possible paths between two end-points in a grid structure comprised of intermediate nodes and multiple links between them. In the problem under study, each node represents a distortion state and each path through all distortion states defines a possible m-subset. The optimal solution is found from the set of all possible paths.

This process of identifying all possible candidates and paths is represented by the trellis shown in Fig. 6.3, where the x-axis represents the stages, t, and the y-axis the states, k, of each stage. Many subsets of cardinality t, correspond to each stage, t, as referred to above. To each state corresponds a candidate segment p_k. Thus, each node in Fig. 6.3 represents a distortion state characterised by the minimum distortion found through all t-subsets ending at p_k, *i.e.*, D_t^k.

The herein described DP algorithm operates in two phases. The first phase consists on constructing the trellis by sequentially defining its nodes and links as functions of (t, k) and identifying a set of segment candidates to each stage. Then a reverse scan of the trellis

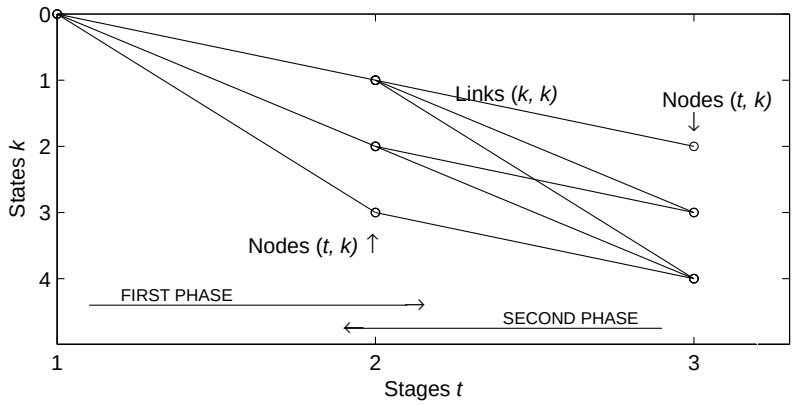


Figure 6.3 – The different candidate solutions dictated from the different pathway combinations

is performed in the second phase, selecting the optimal segment among the candidates at each stage. Next subsection describes this algorithm with further detail.

6.1.3 The Dynamic Programming algorithm

This subsection explains how the Dynamic Programming algorithm works, based on the example presented on page 154, using $n = 5$ and $m = 3$. The two phases explained above are separately described.

PHASE 1

Stage 1

The algorithm begins by putting segment p_0 into the first subset. This is stage $t = 1$ where $\{l_q\} = l_0 = 0$. The 1-subset is thus found.

Stage 2

At stage $t = 2$ a new segment is added among all possible candidates. Note that, as pointed before, not all segments can be candidates at each stage. At stage 2, the possible segments are $k = 1$, 2 or 3. On the one hand, k cannot be less than 1 because the previous one (p_0) is already in the subset. On the other hand k cannot be greater than 3, because at the next stage $t = 3$ it is necessary to have at least one segment available to be added at that stage. Thus, it is clear that in each stage a range interval for k values must be determined, *i.e.*, k_{min}, k_{max}.

In general, when constructing the t-subset, there are $t-1$ segments already used, thus not available. In the extreme case where all $t-1$ segments are contiguous, it means that all indices under $t-1$ are used. Therefore $k_{min} = t-1$. In the case of k_{max}, this is determined by m and current stage, t, such that $m-t$ segments are left available for the $m-t$ futures stages. Since the last segment of U is indexed as $(n-1)$, then $k_{max} = (n-1)-(m-t)$. Thus the possible values for k at stage t are the given by Eq. 6.12,

$$(t-1) \leq k \leq (n-1)-(m-t). \tag{6.12}$$

The bounds imposed by 6.12 dramatically restricts the number of combinations when compared with those identified in Eq. 6.10. The identification of the limits for k determines all the $k_{max} - k_{min} = n - m + 1$ candidate segments to be considered in each stage. This value does not depend on the considered stage; it is constant throughout all stages.

In the current example each stage provides $n - m + 1 = 3$ candidate segments. From 6.12 they are p_1, p_2 and p_3 in the current stage ($t = 2$). This is the reason why three nodes are plotted in Fig. 6.3, in each stage, except in the first one, where there is only one, by definition.

The distortion state, D_t^k, is calculated in each state. This is done by minimising the distortion over all possible sub-subsets $\{l_0, l_1, ..., l_{t-2}, k\}$ that are constructed at node (t, k). Each of these distortion states can be interpreted as being the minimum distortion that is possible to achieve with a subset of t segments that ends in the segment p_k. Then segment p_k is a candidate to consider in the second phase.
According to what has been explained, formally the distortion state is defined as

$$D_t^k = \min_{l_1, l_2, ..., l_{t-2}} \left\{ \sum_{j=0}^{n-1} d\big(p_j, p_i\big) \right\} : i = \max\big\{0, l_1, \cdots, l_{t-2}, k\big\} \wedge i \leq j. \tag{6.13}$$

In Eq. 6.13, D_t^k represents the minimum distortion of a subset with t segments ending in packet p_k, for all the combinations $\{l_1, ..., l_{t-2}\}$. Segments p_0 and p_k do not belong to this optimisation because they are fixed, *a priori*.

An example of the subsets inherent to each state is the following, for a generic 3-subset ending at packet p_3. In this case, the specific subsets $\{p_0, p_1, p_3\}$ and $\{p_0, p_2, p_3\}$ can be

constructed. If packet p_4 is identified as a candidate to be added, then the resulting subsets may be $\{p_0, p_1, p_4\}$, $\{p_0, p_2, p_4\}$ and $\{p_0, p_3, p_4\}$. Thus, each segment p_k defines a series of subsets, over which D_t^k is computed.

Stage t (t=3)

In a general stage t, adding segment $p_{k=l_{t-1}}$ to the $(t-1)$-subset of minimum distortion ending in $p_{l_{t-2}}$ (*i.e.*, $D_{t-1}^{l_{t-2}}$) results in a decrease of the global distortion. Such distortion decrease is given by the difference between the global distortion before and after adding $p_{k=l_{t-1}}$. Since the global distortion in both cases includes a common term corresponding to the set of segments $\{p'_j\} : j \leq l_{t-2}$, then the this difference can be computed by subtracting the distortions obtained from sets $\{p'_j\} : j > l_{t-2}$ and $\{p'_j\} : j > k$. Therefore this differential distortion plays the role of an edge cost between trellis nodes in consecutive stages, $t-1$ and t, *i.e.*, each link in the trellis represents an associated edge cost. Edge costs are represented in Fig. 6.4, generically, by $e^{l_{t-2},k}$.

Concerning the current example, in stage $t = 3$, k is limited to $2 \leq k \leq 4$ leading to distortion states D_3^2, D_3^3 and D_3^4 as shown in Fig. 6.4. Thus, the candidate segments in this stage are p_2, p_3 and p_4. Note that in the previous stage ($t = 2$) several segments p_k (p_1, p_2, p_3) were identified as candidates to belong to the final m-subset, corresponding to the distortion states (nodes) D_2^1, D_2^2 and D_2^3. All the nodes in stage $t = 2$ are linked to other nodes in the current stage $t = 3$ through the corresponding edge costs. For the

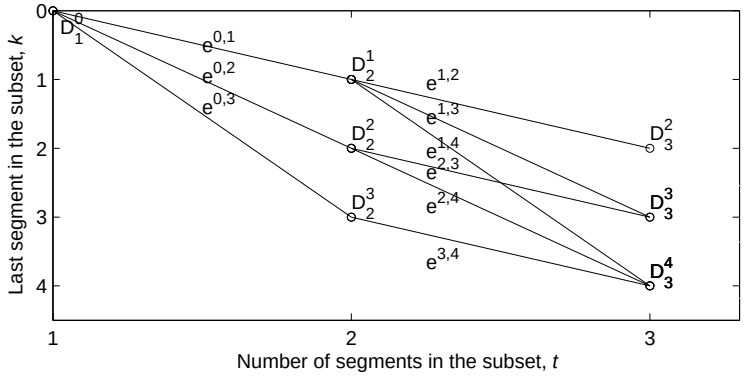


Figure 6.4 – Distortion states and edge costs for the case where $n = 5$ and $m = 3$.

current example ($n = 5$, $m = 3$), the last stage is reached when segments p_2, p_3 and p_4 are

identified as candidates for the last segment of the final 3-subset. Thus, distortion states D_3^2, D_3^3 and D_3^4 are computed, representing the minimum distortions that is possible to achieve among all subsets of three segments ending in p_2, p_3 and p_4, respectively.

Phase 2

The objective of phase 2 of the algorithm is to choose the segments that should be included in the final m-subset, from the candidates previously identified. The selection process starts at the right side of the trellis by choosing the best candidate p_k^*, which is given by

$$p_k^* = p_{l_{m-1}} = \arg\min_k \left\{ D_m^k \right\}. \tag{6.14}$$

Now, moving backwards to stage $t = m - 1$, the question is how to find the best segment to definitively include in the subset, *i.e.*, what is the segment $p_{l_{m-2}}$ that contributes to the minimum overall distortion? The best segment is the one that minimises the difference between distortion states D_{m-1}^k and the edge costs $e^{k,l_{m-1}}$, as given by Eq. 6.15. The underlying idea associated with minimisation of this difference is that a high value of the edge cost corresponds to a high decrease in the global distortion when segment p_k^* is added. On the other hand, a small distortion state means that segment p_k^* will be added to an already small distortion subset. The best segment from stage t, p_k^* is found through the following minimisation,

$$p_k^* = \arg\min_{p_k} \left\{ D_t^k - e^{k,l_t} \right\}. \tag{6.15}$$

Generically, as mentioned before, minimising the above difference corresponds to "maximise the decrease" of the global distortion when the new segment is added to the final m-subset. This selection process is done throughout the trellis and when stage $t = 1$ is reached, the segment p_{l_0} is $p_{l_0=0} = p_0$.

As it can be observed, decisions taken along this second phase take into consideration global information from the past, which was left open for future decision. This knowledge about past and future gives a global set of solutions that permits to derive the minimum distortion one. Thus, it solves the problem the sub-optimal greedy algorithm leaved open. Next subsection presents a mathematical formulation of the algorithm that allows its implementation using a high-level programming language.

6.1.4 Mathematical formulation of the DP algorithm

As described before, the distortion states D_t^k are found through a minimisation process, which can be formulated as:

$$D_t^k = \min_{l_1,l_2,\ldots,l_{t-2}} \left\{ \sum_{j=0}^{n-1} d\Big(p_j, p_{i=\max(l):\, l\in\{0,l_1,\cdots,l_{t-2},k\}\wedge i\le j}\Big) \right\}. \tag{6.16}$$

$l_0 = 0$ and $k = l_{t-1}$ do not belong to the optimisation because p_0 always belong to the subset and, by definition of D_t^k, p_k always belongs, too.

Since p_k is the last packet in a subset, it means that distortion caused by packets beyond p_k, has the constant form $\sum_{j=k}^{n-1} d(p_j, p_k)$, and Eq. 6.16 may be rewritten as

$$D_t^k = \min_{l_1,l_2,\ldots,l_{t-2}} \left\{ \sum_{j=0}^{k-1} d\Big(p_j, p_{i=\max(l):\, l\in\{0,l_1,l_{t-2}\}\wedge i\le j}\Big) + \sum_{j=k}^{n-1} d\Big(p_j, p_k\Big) \right\}. \tag{6.17}$$

In order to reach the recursive format required by DP, Eq. 6.17 can be rewritten by adding and subtracting the same value, $\sum_{j=k}^{n-1} d(p_j, p_{i=\max(l):\, l\in\{0,l_1,l_{t-2}\}\wedge i\le j})$. That is,

$$\begin{aligned} D_t^k = \min_{l_1,l_2,\ldots,l_{t-2}} \Bigg\{ & \sum_{j=0}^{k-1} d\Big(p_j, p_{i=\max(l):\, l\in\{0,l_1,l_{t-2}\}\wedge i\le j}\Big) \\ & + \sum_{j=k}^{n-1} d\Big(p_j, p_{i=\max(l):\, l\in\{0,l_1,l_{t-2}\}\wedge i\le j}\Big) \\ & - \sum_{j=k}^{n-1} d\Big(p_j, p_{i=\max(l):\, l\in\{0,l_1,l_{t-2}\}\wedge i\le j}\Big) \\ & + \sum_{j=k}^{n-1} d\Big(p_j, p_k\Big) \Bigg\}. \end{aligned} \tag{6.18}$$

Since in $\sum_{j=k}^{n-1} d\Big(p_j, p_{i=\max(l):\, l\in\{0,l_1,l_{t-2}\}\wedge i\le j}\Big)$, all i are $i < k$, (since $l_{t-2} < k$), then $\sum_{j=k}^{n-1} d\Big(p_j, p_{i=\max(l):\, l\in\{0,l_1,l_{t-2}\}\wedge i\le j}\Big) = \sum_{j=k}^{n-1} d(p_j, p_{l_{t-2}})$ and Eq. 6.18 can be rewritten as

$$\begin{aligned} D_t^k = \min_{l_1,l_2,\ldots,l_{t-2}} \Bigg\{ & \sum_{j=0}^{n-1} d\Big(p_j, p_{i=\max(l):\, l\in\{0,l_1,l_{t-2}\}\wedge i\le j}\Big) \\ & \underbrace{- \sum_{j=k}^{n-1} \Big[d\Big(p_j, p_{l_{t-2}}\Big) - d\Big(p_j, p_k\Big) \Big]}_{e^{l_{t-2},k}} \Bigg\}. \end{aligned} \tag{6.19}$$

Notice the first term of the second sum of Eq. 6.19 corresponds to subtract the distortions caused by missing segments above or equal to p_k when $p_{l_{t-2}}$ is the last one in the subset and the second term of this sum corresponds to add the distortions caused by missing segments above or equal to p_k, when p_k is the last segment in the subset. Subtracting the whole second sum, $\sum_{j=k}^{n-1}\left[d(p_j,p_{l_{t-2}})-d(p_j,p_k)\right]$, corresponds to the distortion reduction due to the fact of adding segment p_k to the subset when last one was $p_{l_{t-2}}$, *i.e.*, it represents the edge-cost, $e^{l_{t-2},k}$.

By inspection it can be observed that the first sum of Eq. 6.19 has the form of a distortion state of a subset with $t-1$ segments ending in the segment $p_{l_{t-2}}$, thus Eq. 6.19 can be rewritten as

$$\begin{aligned} D_t^k &= \min_{l_{t-2}}\left\{\min_{l_1,l_2,\ldots,l_{t-3}}\left\{\sum_{j=0}^{n-1} d\big(p_j,p_{i=\max(l)\,:\,l\in\{0,l_1,l_{t-2}\}\wedge i\leq j}\big)\right\}-e^{l_{t-2},k}\right\} \\ &= \min_{l_{t-2}}\left\{D_{t-1}^{l_{t-2}}-e^{l_{t-2},k}\right\}. \end{aligned} \tag{6.20}$$

Eq. 6.20 expresses the recursion formula that is necessary for using DP on calculating the needed distortions and edge costs in phase 1.

With this knowledge and the *modus operandi* described above it is possible to establish the expression that serves as the base to computationally derive the indices of the optimal m-subset, $L^* = \{l_t\} = \{l_0, l_1, \cdots, l_{m-1}\}^*$ in phase 2, as expressed in Eq. 6.21. From what has been said, L^* contains the indices that determine which segments of $\{p_i\}$ constitute the payload of the high priority packets.

$$l_t = \begin{cases} \arg\min_k \left\{D_t^k\right\}, & t = m-1 \\ \arg\min_{l_t} \left\{D_{t+1}^{l_t} - e^{l_t,l_{t+1}}\right\}, & t\in[1,m-2] \\ 0, & t=0 \end{cases} \tag{6.21}$$

The classification algorithm described so far is represented in Appendix B.

6.2 Packet loss models

In order to test how efficient the classification algorithm is in determining which segments contribute more to the perceived quality, packet losses simulating tests were carried out in which low priority packets were preferentially lost. Furthermore, two widely used discrete models for packet loss were used to compare with those relative to low priority packet losses. Bernoulli and Gilbert models were used [170–174].

Packet losses are assumed to have two possible causes. On the one hand, network congestion makes routers to discard packets when packets arrive at a faster rate than the link capacity. This kind of loss is referred to as congestive loss [175] and tend to be bursty [176]. There is a correlation between losses. On the other hand, high level of statistical multiplexing, noise and damages contribute to bit error rate. This kind of loss is referred to as isolated or random loss [175, 176]. There is no correlation between this kind of losses. Despite nowadays isolated losses are rare in VoIP, under certain conditions they may occur [177].

Next two subsections are devoted to random and burst statistical models respectively. Bernoulli and Gilbert models are described [170].

6.2.1 The Discrete Random Bernoulli Model

In probability theory and statistics, the Bernoulli distribution is a discrete probability distribution in which only one out of two values may occur. A random variable X with this distribution may only assume the different values $X = 0$ and $X = 1$. It takes the value 1 with probability $Pr(X = 1) = p$. Since the two allowed values are mutually exclusive, the probability to assume the value 0 is $Pr(X = 0) = 1 - Pr(X = 1) = 1 - p$. Let us define $Pr(X = 0) = q$. p and q probabilities are referred to as success and failure probabilities, respectively. In the packet loss context, $X = 1$ with probability p represents a packet loss event.

The probability mass function of this distribution is given by $f(k; p) = p^k(1 - p)^{(1-k)}$ for

$k \in \{0, 1\}$ which can also be more clearly expressed by [178]

$$f(k; p) = \begin{cases} p, & \text{if } k = 1 \\ 1 - p, & \text{if } k = 0 \end{cases} . \tag{6.22}$$

It is important to retain that there is no correlation between successive events, so the events are "purely random".

6.2.2 The Gilbert Model

According to the Federal Standard 1037C, in the communications domain, "burst error" is a contiguous sequence of received symbols that are in error and there exists no contiguous subsequence of s correctly received symbols within the error burst. The integer parameter s is referred to as the guard band of the error burst. In other words, the first symbol in a burst is separated from the last one from the last burst by, at least s symbols [179]. Applying to our case, symbols represent voice packets.

Since packet loss constitutes the main source of QoE degradation of VoIP calls [131, 180] and this is more pronounced if such losses occur in burst [181], which is most the reality in VoIP communications [182], the impact of packet losses on the voice quality must be measured, which can be achieved by using traces of traffic with such errors. In addition, when it is more appropriate, stochastic models can also be used to generate similar error patterns with similar statistic characteristics as the observed signals in real situations. Previous works show that simple Markov models are appropriate to capture the observed loss pattern [172]. In this context, the simple Gilbert model, a particular case of a two-state discrete-time Markov model, is often used to model how VoIP packet losses occur [171, 172, 174, 181]. One advantage of this model when compared with the Bernoulli model is that it considers packet losses as time-dependent events as in real situations. Furthermore, it permits to parameterise the loss burst length.

In order to understand the principles of the Gilbert model, consider Fig. 6.5, where two independent states and four probabilities are defined: two of these probabilities are conditional and two are unconditional. Two independent states are defined to represent two situations: packet loss, often represented by the "one" state and packet reception (no

loss), often represented by the "zero" state[4].

As conditional probabilities it defines

i) the probability of loosing a packet, given that the previous one was not lost, herein denoted by p;

ii) the probability of receiving a packet, given that the previous one was lost, herein represented by q.

Therefore, $1-p$ represents the probability of no loss given that, in the previous event there were no loss, too, and $1-q$ represents the probability of loss given that in the previous event, a loss occurred. These probabilities are also called transition probabilities and they correspond to statistically dependent events.

As unconditional probabilities the model defines the probability of being in state "zero" and the probability of being in state "one", regardless of the previous states. They are, respectively, represented by π_0 and π_1 [183]. In a packet chain, π_1 represents the packet loss probability over the time. Thus, it is a packet loss rate. From conditional probabilities it

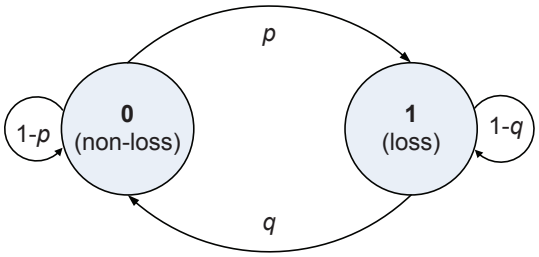


Figure 6.5 – Simple Gilbert Model

is possible to derive unconditional probabilities that are useful in practical tests. In order to clarify the relationship between such parameters at steady state, let us consider the transition matrix, P, and the state probabilities vector, π, both represented in Eq. 6.23.

$$P = \begin{bmatrix} (1-p) & q \\ p & (1-q) \end{bmatrix}, \quad \pi = \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix}. \tag{6.23}$$

[4]The model is meant to represent failures and not successes.

In order to calculate the unconditional probability of being in each state, Eq. 6.24 is used [183],

$$\begin{bmatrix} (1-p) & q \\ p & (1-q) \end{bmatrix} \times \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix} = \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix}. \tag{6.24}$$

From Eq. 6.24 it is possible to isolate π_0,

$$\pi_0 = (1-p)\pi_0 + q\,\pi_1, \tag{6.25}$$

which, together with the fact the sum of the probabilities must be 1, that is,

$$\pi_0 + \pi_1 = 1, \tag{6.26}$$

makes possible to derive the unconditional probabilities π_0 and π_1 of being in state "zero" and state "one", respectively, as given by Eq. 6.27.

$$\pi_0 = \frac{q}{p+q}, \quad \pi_1 = \frac{p}{p+q}. \tag{6.27}$$

In this context, π_0 refers to the percentage of received packets (not lost).

It is possible to compute the probability, p_k, of a burst loss of length k to occur. In [183] it is computed and herein given by Eq. 6.28. Therefore lengths of bursts in this model, have a geometric probability distribution.

$$p_k = P(Y = k) = (1-q)^{k-1}\,q. \tag{6.28}$$

Let us consider Y a random variable that describes the distribution of burst lengths. Since $1-q$ is less than 1, bigger bursts are less probable to occur then small ones.

Based on 6.28 it is possible to compute the *mean burst length*, $E[Y]$. It is given by Eq. 6.29:

$$E[Y] = \sum_{k=1}^{\infty} k\,p_k = \frac{1}{q}. \tag{6.29}$$

As it can be seen, the mean burst length only depends on the loss behaviour of two consecutive packets: the probability of not losing a packet, given that previous one was lost. Typically, higher values of mean burst loss length (lower values of q) correspond to

greater burstiness of missing packets.

It is also possible to compute conditional probabilities, p and q, from unconditional probabilities, π_0 and π_1. It may be useful when analysing traces. One can either inspect the entire trace or using the "loss length distribution statistics" [183]. Let us consider o_i $(i = 1, 2, ..., n-1)$ as denoting the number of loss bursts having length i and $n-1$ the length of the longest burst. o_0 denotes the number of delivered packets. Then, p and q can be calculated by Equations 6.30:

$$p = \sum_{i=1}^{n-1} o_i / o_0; \qquad q = 1 - \left(\sum_{i=2}^{n-1} o_i \, (i-1) \Big/ \sum_{i=1}^{n-1} o_i \, i \right) \tag{6.30}$$

For practical purposes, the packet loss ratio, π_1, and mean burst loss size are important. Based on the presented formulation, needed probabilities were calculated on the tests described in section 6.3.

6.3 Simulated Results from Priority Transmission

This section presents the tests that were carried out to evaluate the performance of the classification algorithm described in section 6.1 to simulate priority VoIP transmission and presents and discuss respective results.

In our experiences we were interested to evaluate the effectiveness of the classification algorithm with VoIP packets according to their importance as referred in section 6.1. The concept of effectiveness herein referred to, refers to the capability of the algorithm to determine the less important segments that, if erased, leads to the minimum voice quality degradation, for a given percentage of less important segments. In other words, this is its capability of determine the segments that represent the original utterance (erasing the remaining ones) with minimal distortion.

Fig. 6.6 illustrates the sequential process used to carry out our experiences. The idea is to classify segments at the send side as either high or low priority such that those of less importance are discarded first in a congested network. Thus, the most important packets are the base of the reconstructed signal at the receive side. Note that not all the less important segments are necessarily lost and not all the important segments are preserved.

It means that losses may occur from just few of the low important segments to, even, some of the most important segments, since real losses are stochastic events.

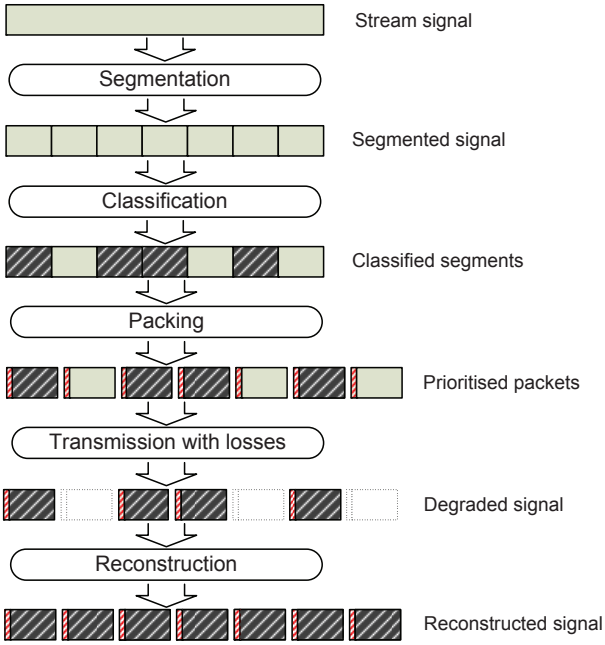


Figure 6.6 – Processes involved in the simulation of transmitted voice with classification

The figure shows how the signal is processed along with the pathway between the send and the receive sides. As it can be seen, the first process consists of segmenting the original voice stream signal. In this process the voice utterances are taken and divided into temporal segments in order to represent the contents of the VoIP packets as in a real VoIP communication. As result, the signal is divided into temporal segments (segmented signal), proper to feed the classification process.

Thereafter, this segmented signal is subject to the classification process to find the optimal m-subset. The m-subset comprises the most important segments, *i.e.*, the payload of the priority packets. We are interested to check whether losing preferentially the low priority packets will lead to the least degradation on voice quality when compared with random-like packet losses.

The next process represents the packet loss occurred in a lossy network. The result is a

signal whose low priority packets are preferentially lost. Light rectangles represent erasures.

The last process consists in the reconstruction of the signal from the high priority packets (reconstructed signal). As it can be seen, erasures are replaced by the last known packets.

To make representative experiments, different voice utterances, with different lengths, spoken by different speakers, as well as different segment sizes and erasure percentages, were used. Each combination of these variables led to different cases. As output results, RMSE between observed and original signals and between reconstructed and original signals were then obtained for each test, as well as MOS derived by using the PESQ algorithm [46].

In this study, four voice utterances extracted from the English sentences recommended by ITU-T Rec. P.501 were used [162]. From the sentence "The juice of lemons makes fine punch", a 400 milliseconds voice sample containing the utterance "punch" was extracted. From the sentence "A large size in stockings is hard to sell", a 400 milliseconds voice sample containing the utterance "A large" was extracted. From the sentence "Glue the sheet to the dark blue background", a 1 200 milliseconds voice sample containing the utterance "background" was extracted. From the sentence "These days a chicken leg is a rare dish", a 1 200 milliseconds voice sample containing the utterance "Is a rare dish" was extracted. The Table 6.1 shows the used utterances and its characteristics. "M" and "F" stand for male and female genders and the followed digit differentiates the speaker.

In these extraction operations, special care was taken in order to include both cases of lower and high energy, fricative and vowel sounds, female and male speakers, and, for each gender, deep and shrill speaker's voice. For all these utterances, downsampling and filtering operations were done in order to represent the 8 kHz voice telephony quality, proper to feed the used ITU-T PESQ application. Furthermore, a silence window was produced by muting 20% and 40% of half utterances in a low energy region.

Our tests were divided into two stages. In the first stage, the aim was to test how effective the classification algorithm is in assigning different levels of importance to segments.

Table 6.1 – Used utterances in the classification process

Utterance	Length (ms)	Speaker	Silence (%)
"Pounch"	400	M1	20
"A large"	400	F1	0
"Background"	1 200	M2	40
"Is a rare dish"	1 200	F2	0

Only the original and degraded signals were used to determine such effectiveness. This is useful to determine how harmful is the packet loss for the voice quality when no voice reconstruction neither any other concealment scheme is used. In the second stage lost signal reconstruction was carried out to test how good the reconstruction is in enhancing the voice quality.

In the first stage, the classification algorithm divided utterances in temporal segments as referred to above. For each utterance, two segment sizes and four percentages of erasures were considered. The segment sizes were defined as 10 ms and 20 ms in order to simulate VoIP packets payload [184]. Erasure percentages of 5%, 10%, 20% and 30% were used to simulate packet loss rates in VoIP communications.

Thereafter the algorithm was run eight times over each segmented utterance to simulate all *segment size* and *erasure percentage* combinations, *i.e.*, $2 \times 4 = 8$ times. Each classification run produced a set of indices that represents which segments are the most important ones. Tanking into account the four utterances, 2 *segment sizes* $\times 4$ *erasure percentages* $\times$ 4 *utterances* = 32 sets of importance indices were thus derived. From other point of view, for each percentage of losses, 2 *segment sizes* $\times$ 4 *utterances* = 8 tests were carried out, which represents 8 results. In this process, each index set is unique for each combination of utterance, segment size and erasure percentage. It is not a stochastic process: multiple runs of the algorithm over the same referred combination, always produce the same result. Notice that in the tests no VoIP codec was used. The voice signal is Pulse Code Modulation (PCM), 8 bits.

In our tests we have considered the extreme case in which 100% of low priority packets are lost and 100% of the high priority packets are preserved. This means that all less important segments were erased and all the most important segments were preserved. To

test the effectiveness of the classification algorithm, the results obtained from losing packets prioritised according to the classification process were compared with those obtained from losing packets according to two random processes: Bernoulli and the Gilbert models. Thus we define three cases for losing packets: loss of low priority packets as "priority loss" and losses dictated by Bernoulli and Gilbert models as "random losses".

In the case of priority loss, the observed signal is reconstructed from the m high priority packets and the $n-m$ erasures. For each pair of original and observed signals, RMSE and MOS values were then obtained and compared.

Concerning the random erasures, for each stochastic model (Bernoulli and Gilbert), 20 random experiences were carried out for each combination of *segment size*, *loss probability* and *utterance*, resulting, respectively, on $20 \times 2 \times 4 \times 4 = 640$ new observed signals. This means that for each percentage of erasures, $20 \times 2 \times 4 = 160$ values were obtained. To be coherent with the erasure percentages used by the classification algorithm, the same percentages were used as loss probabilities, π_1: 5%, 10%, 20% and 30%.

In order to obtain losses according to the Gilbert model, the **sqngen** module from the Linux kernel was run [185, 186]. **sqngen** input arguments included random experiences (20 was chosen) and the mean length error burst (3 was chosen).

In the second stage of our tests, a zero-order hold reconstruction was carried out for the whole universe of the observed signals. This reconstruction consists on substituting each missed segment by the previous known one in the observed signal. As result a new signal (the reconstructed one) was produced from each observed signal and, again, RMSE and MOS were derived –now, between original and reconstructed signals. Results of both stages, represented by means of the achieved RMSE and MOS values, are depicted in Figs. 6.7 to 6.13.

First stage – Degraded *versus* original signals

Fig. 6.7 and Fig. 6.8 show the accuracy of observed signals in comparison with the respective originals as a function of the percentage of missing segments for all used utterances and methods of losing packets. In the former, this is measured by RMSE representing the

distortion between signals. In the latter one this is measured by means of MOS_LQO to represent the subjective voice quality. In both figures all marked points represent mean values. In the case of those relative to the prioritisation model, each mean value is the average of the eight values derived from the 2×4 = 8 (number of utterances×segment sizes) combinations that were formed from all utterances. In the case of the random models, each mean value is the average of $2 \times 4 \times 20 = 160$ values derived from the same eight combinations, where 20 random experiences were carried out for each one, as referred to above. Next to each marked value a vertical bar around each point represents the standard error of that mean[5]. Since this parameter gives a measure of the degree of scattering of the values and these bars do not overlap those of the other lines, these results can be accepted as statistically representative of the whole set of values. Fig. 6.8 also includes a reference line to signal the case in which PESQ does not detect voice degradation (label "MOS Ref").

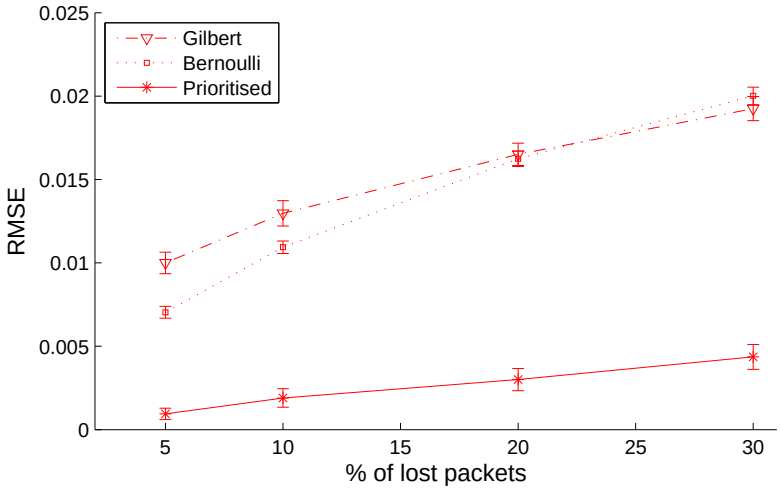


Figure 6.7 – Average distortions between observed and original signals

In Fig. 6.7 it is possible to see that:

i) RMSE values monotonically grow with the percentage of lost packets.

[5]The Standard Error of the Mean, σ_M is given by $\sigma_M = \sigma/\sqrt{N}$, where σ represents the standard deviation of the original distribution and N is the sample size (the number of scores each mean is based upon). Since σ_M divides the standard error by the sample size, it gives a representativeness accuracy of the standard error (the larger the sample size, the smaller the standard error of the mean).

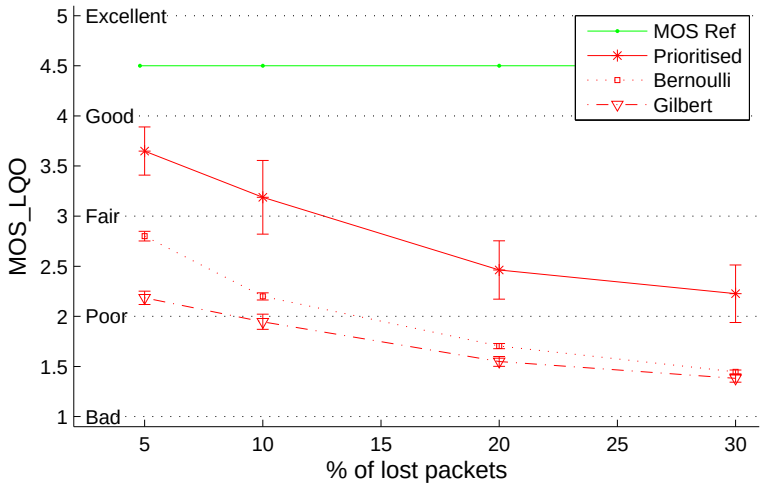


Figure 6.8 – Average voice quality of observed signals using the original as reference

ii) RMSE values obtained from signals that were corrupted according to the random processes are always greater than those obtained from signals whose low priority packets were preferentially lost (100%). Since RMSE measures the distortions of the processed signals in comparison with the originals, the most important fact to retain from these results is that the signals whose low priority packets were preferentially lost are always less distorted than those ones randomly corrupted. In other words, when applying this classification algorithm to VoIP, signals formed by the high priority packets are expected to be lesser distorted. Therefore, it is expected that users experience better voice quality when hearing respective utterances.

As expected, higher percentage of lost packets corresponds to higher signal degradation for all packet loss models.

Fig. 6.8 shows the results obtained from the same experiments but the voice quality is measured by the Mean Opinion Scores (MOS_LQO). Contrary to RMSE, high MOS values are better values since they are proportional to the perceived voice quality. Higher values correspond to lower distortions. In this figure it is possible to see that:

i) MOS values monotonically decrease with the percentage of missing segments.

ii) MOS values obtained from prioritised signals are always greater than those obtained from utterances randomly corrupted. Since MOS measures the quality of processed

utterances in comparison with their originals, the most important fact to retain from these results is that losing the segments classified as less important always exhibit better voice quality than losing segments according to either Bernoulli or Gilbert models, for all percentages of missing segments.

Also important is that quality of utterances in which packets were randomly lost is not recommended, since MOS scores are always below 3 and even below 2, to which correspond "Poor" and "Bad" opinions, according to ITU-T Rec. P.800, *i.e.*, "many users dissatisfied" and "nearly all users dissatisfied", according to ITU-T Rec. G.109 [118, 142], visible in Table 3.12. On the contrary, for utterances formed by the prioritised packets, the average of MOS scores is equal to or greater than 3 for 5% and 10% of lost packets (see line "Prioritised"). According to the ITU-T Rec. P.800, this value corresponds to an opinion of "Fair"; thus acceptable.

In both Fig. 6.7 and Fig. 6.8 it can also be seen that standard error of the mean values relative to the mean values concerning the prioritisation process do not overlap those ones concerning the random processes. This means that the probability to have cases in which RMSE or MOS obtained from prioritisation can attain MOS values obtained from random processes is very small, even null, since most of the values are inside the bar and their extremities are very far from each other. These values also show how confident and representative the plotted means are on representing the population of all possible random experiences.

Table 6.2 shows the values plotted in Fig. 6.8 and expresses the MOS differences between observed signals that were corrupted according to the prioritisation model and the observed signals that were corrupted according to the random methods. Values in the row labeled as "Random (avg)" are an average of the Bernoulli and the Gilbert models values. As it can be seen, the perceptual quality of the observed prioritised signals is greater than that of the observed signals that were randomly corrupted. In fact the average differential MOS is $\Delta MOS = 0.9802$. As a whole, these results corroborate those of Fig. 6.7. signals formed by high priority packets make users to experience better voice quality than those that were randomly corrupted. From these results it is possible to see that the proposed

classification algorithm is very efficient when applied to voice signals.

Table 6.2 – MOS differences between prioritised and randomly corrupted signals using the original as reference

	5%	**10%**	**20%**	**30%**	**Average**
Prioritised	3.6495	3.1879	2.4633	2.2260	2.8817
Random (avg)	2.4923	2.0728	1.6270	1.4138	1.9015
Differential, ΔMOS	1.1572	1.1151	0.8364	0.8122	**0.9802**

Second stage – Reconstructed *versus* original signals

Fig. 6.9 and Fig. 6.10 show the accuracy of reconstructed signals in comparison with the original signals as a function of percentage of erased segments for all the utterances and methods of losing packets. The same previously used notation is used here to represent the obtained values and used methods (labels, lines, symbols, ...). Fig. 6.9 shows distortion

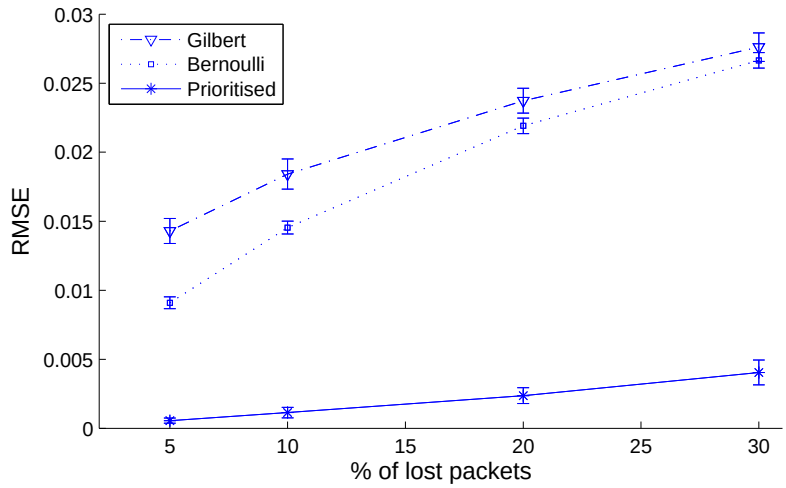


Figure 6.9 – Average RMSE between reconstructed and original signals

after reconstruction of observed signals. As it can be confirmed, the distortion monotonically grows with percentage of lost packets for both the prioritisation and random loss processes and the distortion of the former is always much lower than the latter. For lower percentages of lost packets this relation is more than one order of magnitude lower. The standard error of the mean values remains low, which ensures a good confidence level of the data. Again, it is expected that users experience better quality when listening

to reconstructed signals from the prioritised stream rather than from the ones randomly corrupted. The best results were consistently obtained when the signals are able to keep the most important packets in lossy networks through prioritisation.

Results shown in Fig. 6.10 were obtained from the same experiments as those shown in Fig. 6.9 but they represent the voice quality measured as MOS. As it can be seen again, the voice quality monotonically decreases as the percentage of missing segments increases. This is valid for both priority and random loss. In Fig. 6.10 it is also possible to see that the voice quality of signals reconstructed from priority loss is always greater than the voice quality achieved by signals reconstructed from random losses. Except in the case of 30% of losses, in all cases the MOS of reconstructed signal represent opinions ranging from almost 3 ("Fair") to 4 ("Good") which means utterances with acceptable voice quality.

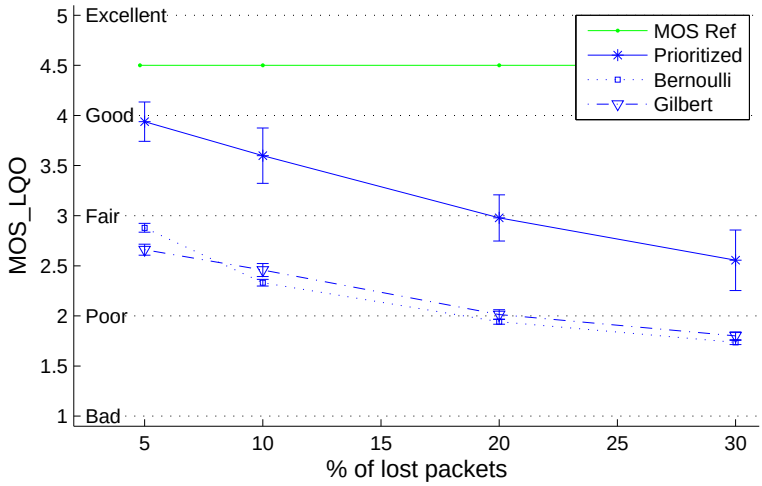


Figure 6.10 – Average MOS between reconstructed and original signals

These results and those shown in Fig. 6.7 and Fig. 6.8 permit to conclude that the classification algorithm is well appropriated in both cases:

i) if an error concealment technique is not feasible, prior classification and prioritisation of voice packets leads to the best results;

ii) if reconstruction is possible, the signal obtained from the high priority packets lead to better results then those randomly corrupted.

Reconstructed *versus* degraded signals

Fig. 6.11 and Fig. 6.12 explicitly show the enhancements achieved by performing the zero-order hold reconstruction over observed signals. In both figures the same conventions as in Figs. 6.7 to 6.12 about data representation apply, except that lines marked with asterisks (label "Observed") refer now to observed signals and lines marked with dots (label "Reconstructed") refer to reconstructed signals. The scale of Figs. 6.7 to 6.10 was intentionally preserved in order to better compare results.

Fig. 6.11 shows the distortions of both observed and reconstructed signals as a function of the percentage of lost packets by means of RMSE values. Apart from both observed and reconstructed lines are monotonically growing, it can also be seen that values concerning reconstructed signals are always lower than values concerning observed signals. This means that reconstructed signals are less distorted than the observed ones, so enhancements on the subjective voice quality are expected to be achieved.

Fig. 6.12 shows the voice quality of both observed and reconstructed signals as a function of the percentage of lost packets by means of MOS values. Again, both functions present a monotonically decreasing behaviour. As it can be seen, voice quality of reconstructed signals is significantly greater than the voice quality of observed signals, despite the fact standard error of the mean values overlaps. Table 6.3 shows the exact values plotted in Fig. 6.12 as well as the enhancements caused by reconstruction. In the universe of all the percentages of losses, reconstruction brought up the scores by an average value of 0.3775.

Fig. 6.12 also shows that, even without reconstruction, MOS values concerning 5% and 10% of lost packets are greater than 3 ("Fair") for which it is expected to not have more than "Many users dissatisfied". In the particular case of 5% of missing segments, the quality attained the value $MOS \lesssim 4$ which corresponds to the opinion "Good" and expected to have only "Some users dissatisfied". However, for percentages above or equal to 20%, MOS values are $MOS < 3$ to which corresponds the opinion "Poor" which is not recommended since it would lead to "Nearly al users dissatisfied".

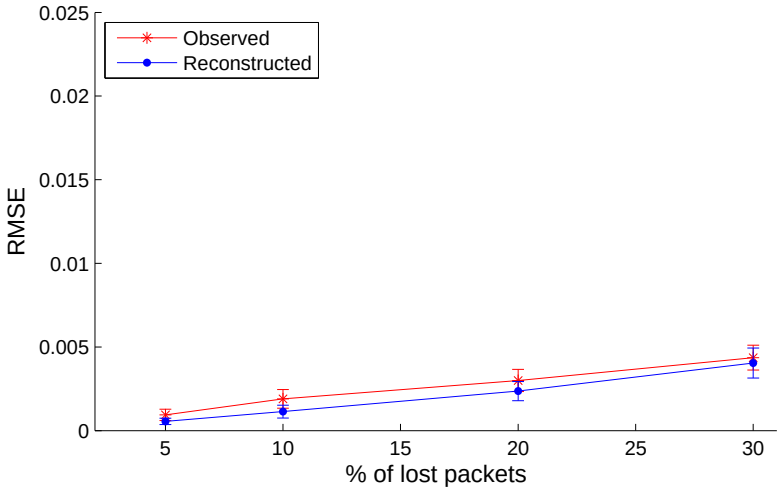


Figure 6.11 – RMSE enhancements attained with reconstruction

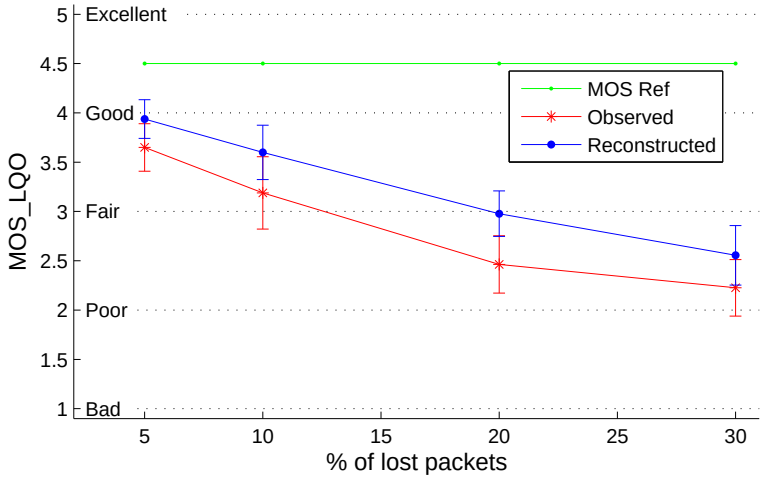


Figure 6.12 – MOS enhancements attained with reconstruction

Also important to notice is that in the case of 20% of missing packets, reconstruction brought up the opinion score from $MOS \lesssim 2.5$ to $MOS \lesssim 3$. This value is near the same obtained by the observed signal when 12.5% packets are lost (3.1879). This means that, in the present case, the proposed classification and reconstruction makes possible to loose near 60% more packets and still keep the same voice quality, than it would happen in the case of pure random losses. It suggests that the use of a more efficient reconstruction

technique can lead to better results, specially if it is specifically taken into account during the classification process.

Table 6.3 – MOS enhancements achieved by reconstruction applied to classified observed signal

	5%	**10%**	**20%**	**30%**	**Average**
Reconstructed	3.9381	3.5989	2.9777	2.5556	3.2676
Observed	3.6495	3.1879	2.4633	2.2600	2.8902
Enhancements	0.2886	0.4110	0.5144	0.2960	**0.3775**

Finally, Fig. 6.13 shows the overall average of the obtained MOS corresponding to all tests (including all voice utterances, all percentages of losses, two segment sizes and all random realisations). Each bar corresponds to a lossy method, as depicted. The lower side of each bar represents the MOS of observed signals. Stacked on them are represented the voice quality increases achieved with the used reconstruction. As it can be seen, while the random losses led to poor voice quality in both random methods, losses occurred according to the prioritisation model led to near-"Fair" voice quality. Furthermore, proceeding with reconstruction, this is also the method that gives the best results, attaining the "Fair" voice quality. Remember that these averages include the cases of 20% and 30% of packet losses, which are not usual in real voice communications. It can also be seen that even signals that were corrupted according to the prioritisation model are better then the reconstructed ones from when random corruption was taken place.

Overall, results exhibited in Fig. 6.13 answer the initial question about the effectiveness of the proposed classification algorithm. Since for both situations of corrupted and reconstructed signals, MOS values relative to prioritised packets are better than those relative to the randomly corrupted signals, we can conclude that the proposed algorithm is appropriated to classify VoIP packets as high and low priority according to their individual relevance to the global perceptual voice quality.

In this chapter, so far, the concept of prioritising packets according to their importance to the voice quality has been proved as a valuable contribute to enhance the voice quality in networks with the ability to prioritise packets. The used reconstruction method is the zero-order hold interpolation. In chapter 4, linear interpolation methods were used to

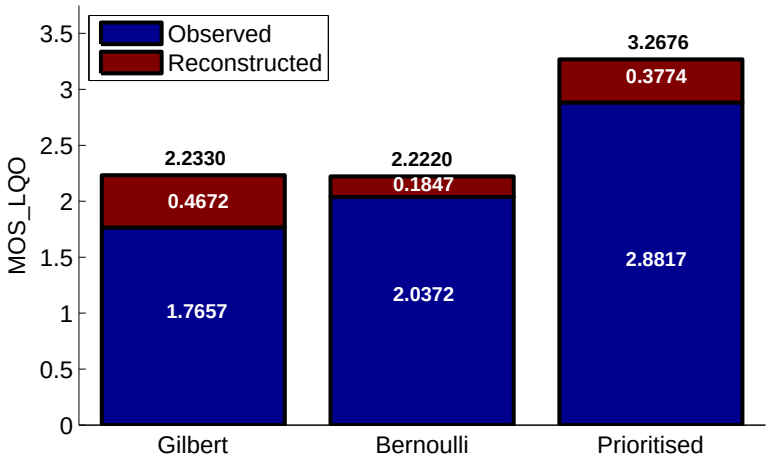


Figure 6.13 – MOS enhancements attained with classification-and-prioritisation without and with reconstruction for random and prioritisation models of losing packets

also perform signal reconstruction. In the next section simulated tests explore a combined technique formed by the packet classification followed by the zero-order interpolation and a maximum dimension linear interpolation algorithm as a way to lever the performance of the signal reconstruction task.

6.4 Packet prioritisation combined with Papoulis Gerchberg algorithm

This section presents an investigation to study the contribution of packet prioritisation followed by zero-order hold interpolation to the maximum dimension linear interpolation algorithm presented in cap. 4: the Papoulis-Gerchberg (PG) algorithm. The underlying idea is to combine both interpolation techniques in which the output of the zero-order hold interpolation of a prioritised lossy signal is the input of the PG algorithm, *i.e.* the observed signal referred to in section 4.2. From this strategy results a combined reconstruction method in which the zero-order hold interpolation of the least important packets is used as the first iteration of this combined method.

With this study we were interested to determine if this strategy can reduce the number of iterations of the PG algorithm in comparison to the case where it is used alone to reconstruct the signal. From another point of view, we were interested to determine if, for a certain number of iterations, it is possible achieve better error values between the observed and the reconstructed signals.

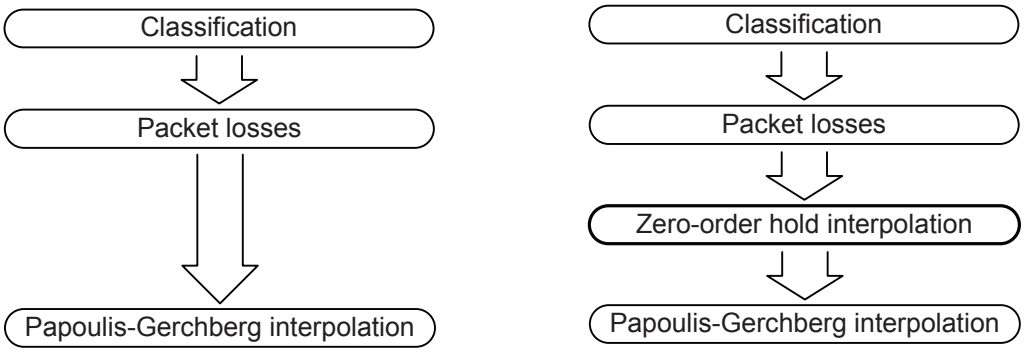
In our tests, the referred strategy included four main processes:

- Packet classification in high and low priority,
- Packet losses,
- Zero-order hold reconstruction and
- Papoulis-Gerchberg reconstruction,

as described so far in the current chapter and in chapter 4, respectively.

Since we were interested to measure the contribute of packet prioritisation followed by zero-order hold reconstruction to the PG reconstruction, the simulation tests were carried out with and without zero-order interpolation. Fig. 6.14 shows a comparative synopsis of the processing chains involved in each case.

In the tests, three packet sizes were used containing two, three and four samples, re-



(a) Papoulis-Gerchberg reconstruction

(b) Zero-order hold followed by Papoulis-Gerchberg reconstruction

Figure 6.14 – The use of the zero-order hold interpolation as an aid to the Papoulis-Gerchberg algorithm

spectively. In case of packet losses, these sizes represent burst lengths in the signal to reconstruct. Concerning the PG reconstruction, two stop criteria were used: the residual error between the signals obtained in consecutive iterations and the error between the reconstructed signal and the observed (corrupted) one, both measured as the RMSE. Two different oversampling factors were also used: $r = 0.6$ and $r = 0.4$, as it was in chapter 4.

The simulation tests aimed to evaluate and compare the reconstruction performances in both cases, *i.e.*, with and without zero-order hold interpolation. Two sets of tests were defined, according to the performance metric used for comparison performances.

First set of tests

In the first set, the number of iterations needed to attain a certain target error was measured. In the second set, the error was measured, for a predefined number of iterations.

Concerning the first set, the results are depicted in Figs 6.15 to 6.17. Fig. 6.15 shows the number of iterations needed to reach a residual error of 10^{-8} for an oversampling factor, $r = 0.6$. As it can be seen, the use of a signal reconstructed by the zero-order hold interpolation as input to PG algorithm always save iterations: for burst lengths of 2, 3 and 4, the number of iterations decreased from 110 to 76, from 659 to 514 and from 3883 to 2633, respectively. For the same amount of error, starting the PG algorithm with zero-order interpolated samples rather than zeros, reduces the total number of iterations by 31%, 22% and 32%, respectively.

Fig. 6.16 shows the number of iterations needed to reach an error of 3×10^{-4} between reconstructed and observed signals for an oversampling factor, $r = 0.6$. Similar to the previous case, there is also a decrease in the number of iterations from 33 to 7, 286 to 141 and from 2028 to 778, respectively. This corresponds to reduce the total number of iterations by 79%, 51% and 62%, respectively.

Fig. 6.17 shows the number of iterations needed to reach an error of 10^{-5} between reconstructed and observed signals for an oversampling factor, $r = 0.4$. Similar to the previous cases, there is always a decrease in the number of iterations: 18 to 14, 47 to 33 and 137

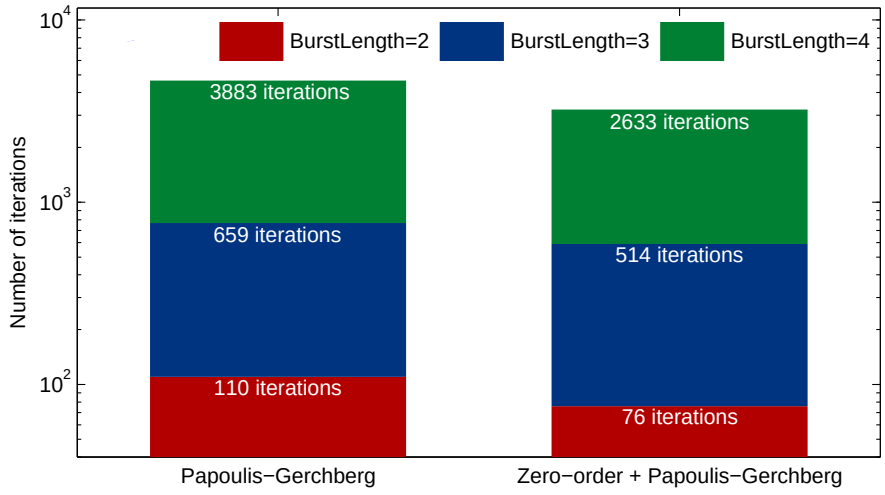


Figure 6.15 – Number of iterations needed to reach a residual error of 10^{-8} (r=0.6).

to 120 for burst lengths of 2, 3 and 4, respectively. Again, this corresponds to a reduction of iterations by 22%, 30% and 12%, respectively.

Overall, from these results it can be seen that the combined method always leads to better performance. Table 6.4 shows the number of iterations needed to reach the specified criteria, for both reconstruction methods, as well as the savings achieved in the number of iterations, which reflects the computational effort. Each row corresponds to a different burst length which is subdivided into three rows representing, each one, values relative to the three cases described in Figs. 6.15, 6.16 and 6.17, respectively. Column "PG" refers to the number of iterations required by the PG algorithm, on its own. Column "ZO+PG" shows the number of iterations when using the combined strategy. Column "Savings" shows the number of iterations saved by the combined technique, that is, the difference between both methods, and "% savings", the respective percentage. "Average" represents the percentage of all saved iterations for respective burst length and "Overall" represents the overall percentage of saved iterations for the universe of the tests carried out. Overall, for our test conditions, 40% of less iterations were needed.

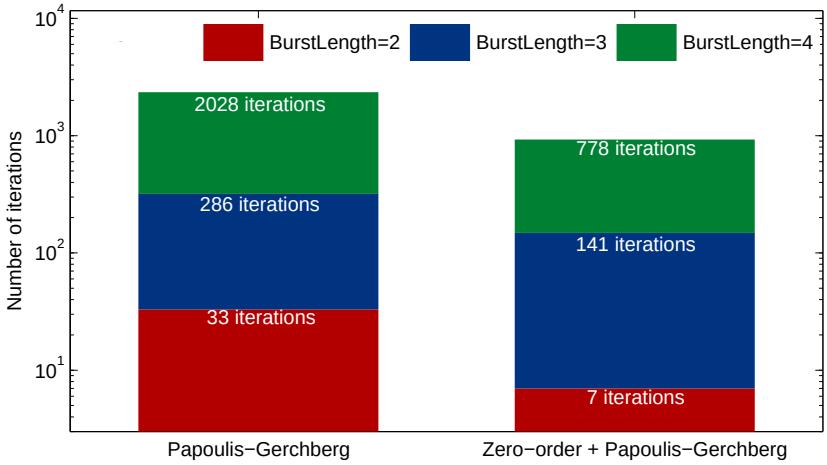


Figure 6.16 – Number of iterations needed to reach an error of 3×10^{-4} (r=0.6).

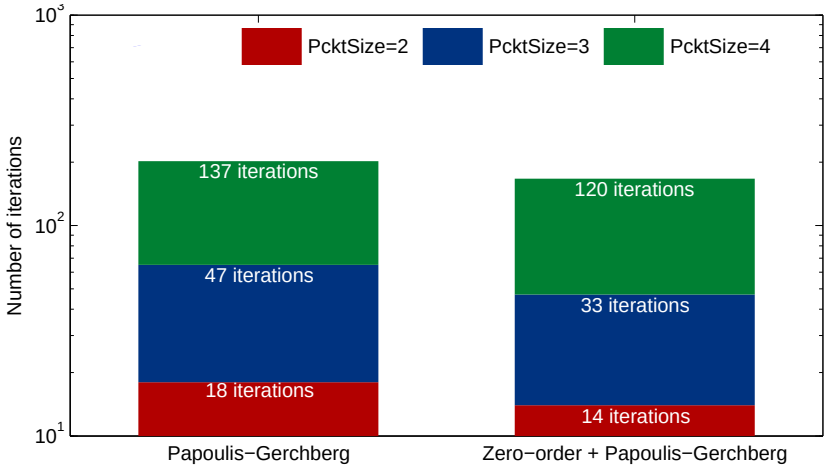


Figure 6.17 – Number of iterations needed to reach an error of 10^{-5} (r=0.4).

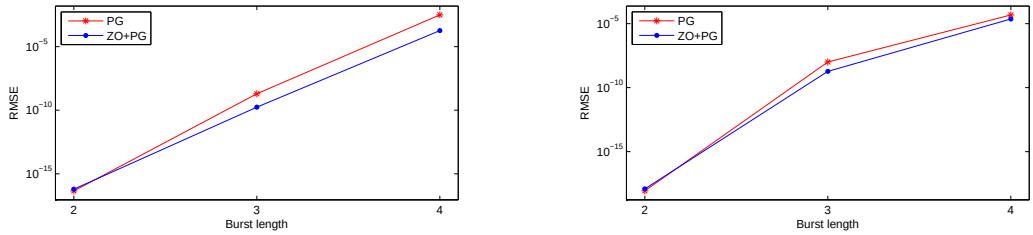
Table 6.4 – Computational effort savings by using zero-order hold (ZO) with Papoulis-Gerchberg (PG) interpolation.

Burst length	PG	ZO+PG	Savings	% savings	Average	Overall
2	110	76	34	31%	40%	40%
	33	7	26	79%		
	18	14	4	22%		
3	659	514	145	22%	30%	
	286	141	145	51%		
	47	33	14	30%		
4	3883	2633	1250	32%	42%	
	2028	778	1250	62%		
	137	120	17	12%		

Second set of tests

The second set of tests were intended to evaluate the enhancements concerning to the error between the reconstructed and the observed signals, given a certain number of iterations. The test conditions are similar to the previous ones, *i.e.*, the same oversampling factors and burst lengths. The results are presented in Figs. 6.18 to 6.21.

Fig. 6.18 shows the error as a function of the burst length for a fixed number of iterations, for the two methods under comparison. Sub-figure 6.18(a) plots the results achieved for $r = 0.6$, whereas sub-figure 6.18(b) plot the results achieved for $r = 0.4$. Since this factor affects the convergence rate, as studied in chapter 4, a lower number of iterations were used for $r = 0.4$.



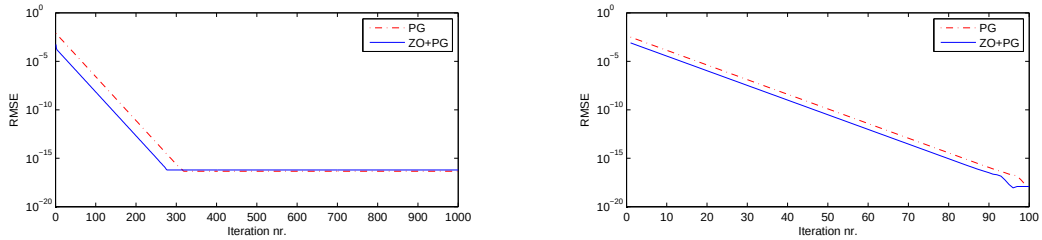
(a) $r = 0.6$, 1000 iterations (b) $r = 0.4$, 100 iterations

Figure 6.18 – Achieved errors for a given number of iterations

As it can be seen, the RMSE values obtained for the combined method (label "ZO+PG") are better than those obtained for PG, only (label "PG"). It can also be seen than this enhancement has more relevance for larger burst lengths, where the PG algorithm tends to exhibit worst performance.

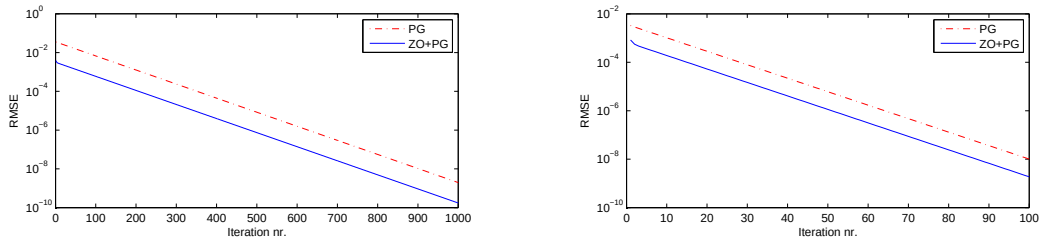
Figs. 6.19 to 6.21 show how the reconstruction error evolves with the iterative reconstruction processes. These figures show that, in general the error values obtained from the combined technique ("ZO+PG") are better than those from Papoulis-Gerchberg technique ("PG"). The only exceptions concern the case in which the burst length is 2 (Fig. 6.19) where, beyond the $\approx 250^{\text{th}}$ iteration (for $r = 0.6$) and the $\approx 90^{\text{th}}$ iteration (for $r = 0.4$), the error values of the combined technique are slightly higher. These exceptions are not considered as relevant since they occur late in the reconstruction, which means that by an appropriate decision about when the process must be stopped, these discrepancies can

be avoided. Furthermore, the difference between these values is very small when compared to the corresponding differences that exist in the beginning of the reconstruction. Since processing time may be a constraint and accurate reconstruction is almost achieved in the first iterations, many iterations are not expected to be necessary and so, the referred exception will not occur. This figure also explains the slight discrepancy showed in Fig. 6.18 in which RMSE from combined reconstruction are slightly worse for burst lengths of 2. This is because the RMSE values were collected precisely at points (iterations) beyond which the error lines invert relative positions. Aside this exception, the combined reconstruction technique always leads to better results.



(a) $r = 0.6$, 1000 iterations (b) $r = 0.4$, 100 iterations

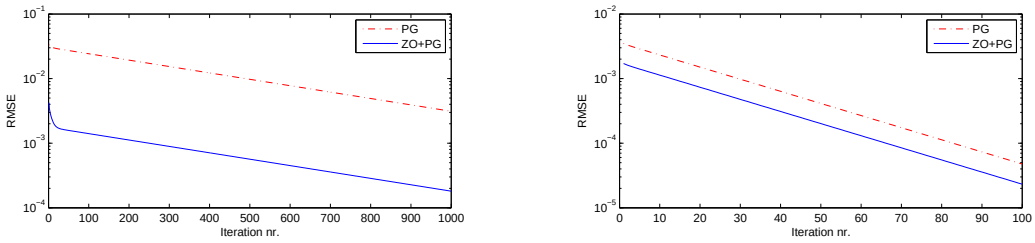
Figure 6.19 – Evolution of the reconstruction error for a burst of length 2



(a) $r = 0.6$, 1000 iterations (b) $r = 0.4$, 100 iterations

Figure 6.20 – Evolution of the reconstruction error for a burst of length 3

Some of the RMSE values shown in the graphics are explicit in Table 6.5. This table shows the error values obtained at the beginning and at the end of both reconstruction methods (PG and ZO+PG) for the different burst lengths and oversampling factors, so that an overall value of error gain can be calculated. Each row corresponds to a different burst length which is, in turn, subdivided into two rows representing, each one, the values relative to $r = 0.6$ and $r = 0.4$, respectively. Columns represent the values obtained in



(a) $r = 0.6$, 1000 iterations

(b) $r = 0.4$, 100 iterations

Figure 6.21 – Evolution of the reconstruction error for a burst of length 4

the first and last iteration from either the Papoulis-Gerchberg algorithm or the combined method (labels "PG" and "ZO+PG", respectively). The table also shows the ratios between the error from each reconstruction method. For example, considering the burst length of 2 and its first row ($r = 0.6$), it is possible to see that in the first iteration of the Papoulis-Gerchberg reconstruction an error of 7.8×10^{-3} was achieved while the error achieved with the combined technique is 5.5×10^{-4}. The ratio between both values is 14, which means that the error achieved by using the combined method is 14 times lower. The error values of the last iteration are presented in the next columns and the ratio is 39. From these two ratios, an average of 26 is obtained, as shown in the respective column (label "Average"). The same reasoning was applied to the remaining burst lengths and oversampling factors and the respective average values, are also shown. An overall average of 9.9 was computed, as shown in the respective column (label "Overall").

Overall, for all tests, it was possible to determine that associating the classification followed by zero-order hold interpolation to the Papoulis-Gerchberg algorithm permitted to decrease the reconstruction error by a factor of $\lesssim 10$.

In this section the gain that is possible to achieve by using the packet prioritisation followed by zero-order hold interpolation combined with the Papoulis-Gerchberg algorithm was explored. It was found that computational effort (measured by the number of iterations) and reconstruction error can be decreased. However, this kind of tests can go further, namely making them more representative in what concerns burst lengths or used signals as well as exploring timing issues, having a practical implementation in mind.

Table 6.5 – Error values (RMSE) at beginning and end of reconstruction for both reconstruction techniques (Papoulis-Gerchberg alone (PG) and Papoulis-Gerchberg preceded by Zero-Order hold interpolation (ZO+PG))

Burst length	First iteration			Last iteration			Average	Overall
	PG	ZO+PG	Ratio	PG	ZO+PG	Ratio		
2	7.8×10^{-3}	5.5×10^{-4}	14	4.3×10^{-14}	1.1×10^{-15}	39	26	9.9
	3.2×10^{-3}	8.1×10^{-4}	3.9	3.6×10^{-15}	8.8×10^{-16}	4.1	4	
3	3.5×10^{-2}	3.8×10^{-3}	9	2×10^{-9}	1.7×10^{-10}	11	10	
	3.3×10^{-3}	8.3×10^{-4}	4	10^{-8}	1.9×10^{-9}	5	4.5	
4	3.3×10^{-2}	4.2×10^{-3}	7.9	3.1×10^{-3}	1.8×10^{-4}	17.2	12.6	
	3.5×10^{-3}	1.7×10^{-3}	2	4.8×10^{-5}	2.3×10^{-5}	2	2	

Nevertheless, we can conclude that the results are encouraging enough to take the concept as valid and to propose a simple and robust signal reconstruction model as future work.

6.5 Conclusions

In this chapter an algorithm to classify voice segments according to their relevance to the voice quality of a utterance has been proposed. In section 6.1 the algorithm was presented and its *modus operandi* explained. In section 6.2 two statistical distributions that model random packet losses were briefly presented. In section 6.3 a classification algorithm was proposed as the base of a packet loss prioritisation and its performance was studied using the two random models. By comparing results, the effectiveness of the algorithm was validated.

Comparative results show that prioritised signals are less distorted and have better MOS values than signals randomly corrupted. As it was presented, a MOS enhancement of 0.98 was achieved. On the other hand, by applying an appropriated reconstruction technique it is possible to further enhance these signals. In our tests it was possible to enhance the average MOS by about 0.38 for these signals, whereas concerning the randomly corrupted signals this enhancement was only about 0.32[6].

[6]This value is an average value coming from both Gilbert and Bernoulli methods.

In section 6.4 a combined technique formed by the prioritisation and the Papoulis-Gerchberg algorithm was tested. The results show that enhanced reconstruction performance can be achieved when compared with the Papoulis-Gerchberg algorithm alone. The simulation tests revealed that a reduction of 40% of the iterations needed by the PG algorithm can be obtained (when fixing the target reconstruction error between observed and reconstructed signals). Alternatively, a decrease of the reconstruction error by a factor of ≈ 10 can also be achieved (when fixing the target number of iterations).

Overall these results permit to conclude that if the proposed classification algorithm is used to classify voice segments and prioritise packets such that more priority is given to the most perceptually important packets in a priority-enabled network, the overall voice quality will be superior than if packets are lost in the usual random way.

7 Conclusions and future work

In this thesis, contributions to enhance and evaluate the voice quality experienced by telephone users have been proposed. Section 7.1 presents the main conclusions drawn from the presented work whereas section 7.2 presents some aspects this work leaved open that are important to further research.

7.1 Conclusions

In chapter 2 the most significant techniques and methods for enhancing and evaluating telephony voice quality were presented. The problem of the impairment factors that degrade the quality of a speech conversation has been studied and special focus was given to the listen difficulty and so the experienced voice quality. In that perspective this chapter dealt with two faces of the same coin: in section 2.2 the most significant techniques to enhance the VoIP quality existent in the literature were presented. On the one hand, techniques for packet loss concealment and recovering were presented and discussed in subsection 2.2.1. On the other hand, QoS enhancement and packet prioritisation were covered and, in section 2.2.2. Section 2.2.3 refers others techniques to enhance VoIP.

In chapter 3 the most important voice quality evaluation methods with interest for our work were presented. The importance of the voice quality evaluation was established and the most significant voice quality evaluation methods were presented. The differences

between subjective tests and objective tests (to measure the subjective quality) as well a parametric method to evaluate the subjective voice quality was studied and discussed. The most important concepts, procedures and methods that are the basis to carry out with subjective evaluating tests were presented. Conversation-opinion tests, Listening-opinion tests, DCR method, CCR method and the CETVSQ method were described as well as the opinion scales that support these methods. Concerning the objective methods, PAQM, PSQM and PSQM$^+$, PAMS, QVoice, PESQ, the Single-ended method for objective speech quality assessment in narrow-band telephony applications and the recently released P.OLQA methods were described. Special importance is given to the PESQ method since it is widely accepted to evaluate narrow-band telephony voice quality. Concerning the parametric method, the E-Model was described. The input parameters as well as the reference model were described in the perspective to be used in the present work.

In chapter 4 two linear interpolation algorithms to recover missing voice samples in a sequence were investigated. In section 4.1 the most relevant concepts of linear algebra regarding the voice reconstruction methods were described. In section 4.2 the referred algorithms were described. Maximum dimension and minimum dimension concepts relative to the reconstruction algorithms were established, as well as the band limiting operation as the *sine qua non* condition to achieve the needed reconstruction and the base to derive the mathematical formulation of the minimum dimension algorithm. Further important concepts were also defined, such as eigenvalues (and spectral radii) and condition number of the system matrix as the parameters that condition the interpolation problem. Their relationship with oversampling and interleaving factors was explored to conclude that it is possible to put, *a priori*, the reconstruction problem in a well conditioning point. In section 4.3 simulation results coming from using the described algorithms were presented. They cover the study of the influence of the spectral radius of the system matrix in the algorithm convergence. The error geometry as well as the signal bandwidth were also object of the study. Results permitted to conclude that these algorithms are better suited to recover missing sample values when the samples are lost in an interleaved geometry. This fact permits to propose to interleave the voice samples at the time they are packetised as a way to increase the robustness of the original signal and better reconstruct the

corrupted signal by putting the problem in a well-conditioning point that is possible to choose *a priori* by an adequate choice of oversampling and interleaving factors.

In chapter 5 a non-reference model to evaluate the voice quality in both analog and packet switching environments was presented. It is based in the E-Model model and was calibrated using PESQ as reference. It is composed by two modules. The first module was presented in sections 5.2 and 5.3 and is intended to be used in the analog circuit-switching environment where the switches Siemens EWSD and Alcatel Sytem 12 are used, both for local and long distance calling areas. Achieved accuracy goes from $|MOSError| = 0.014$ to $|MOSError| = 0.83$ depending on the used switches. In section 5.4 the second module aiming to evaluate the voice quality in a VoIP environment is derived. It supports the G.711, G.729 and G.723.1 codecs and complies with the class C2 of accuracy specified in the IUT-T Rec. P.564. It was tested and put in production in Portugal Telecom Comunicações.

In chapter 6 a packet classification algorithm to distinguish more important and less important packets in regard to their contribution to the perceptual voice quality of a whole message was proposed. The algorithm is useful to individually assign priorities on routing each voice packet so that less important packets are preferentially discarded in a congested network. Section 6.1 proves that not all voice packets have the same impact on the perceptual quality and describes the *modus operandi* and the mathematical formulation of a dynamic programming algorithm that classifies each packet by having the knowledge of the impact a given packet has on the whole message as well as the relationship between packets on the contribution to this whole quality. Section 6.2 briefly describes two statistical distributions useful to simulate random packet losses used in section 6.3. Section 6.3 presents the simulated results that come from comparing the quality of corrupted voice messages, when respective packets are randomly lost with that of corrupted voice messages when low priority are first lost. These results show that the distortion of the randomly corrupted voice signals is greater than the distortion of the signals that were corrupted according to the prioritisation model. By using MOS, it was possible to measure a global enhancement of about 0.98 when comparing random packet losses with priority-driven packet loss. Furthermore, by applying an appropriated error recovering

technique it has was possible to lever the MOS by about 0.38 in prioritised signals against 0.32 in random losses. In section 6.4 a technique that combines both algorithms of packet classification and Papoulis-Gerchberg (PG) was tested and proposed so that accuracy of reconstruction can be enhanced either by decreasing the reconstruction error or by reducing the computational effort of reconstruction when the PG algorithm is used.

Responding to the intelligibility challenges posed by the VoIP technology and its constant growth as referred in the first chapter, the work of this thesis has shown that it is possible to enhance the voice quality by proposing effective methods to increase the robustness of transmitted signals, to recover lost signals, as well as assessing the voice quality experienced by users.

During the development of the present work, new topics arose that are important to further research. Next section presents some relevant topics for further research.

7.2 Future work

Maximising intelligibility by using the appropriate interleaving factor – In chapter 4 two reconstruction algorithms to recover missing sample values were proposed. These algorithms base their usefulness strategy on the interleaving of the signal samples during packetisation. Although the presented discussion pointed out the possibility to put the reconstruction problem in a well-condition point, the interleaving assumption introduces additional delays on the entire path delay. For example to packetise a signal segment of 20 ms it is required to wait 20 ms for the complete collection of samples has been done. This represents an additional delay of 20 ms in the whole end-to-end path when compared with circuit switching technology. 20 ms of delay is not harmful since it is a negligible value when compared with the acceptable value of a global value of 150 ms [22, 23]. However, if for example an interleaving factor of 1:5 is used to combine with a given oversampling factor (let us say r = 0.6 to cite the example illustrated in Fig. 4.8) in a way the problem can be put in a well-conditioned point, it is required to wait for the first sample of the 5^{th} segment of the original sequence to complete and release the first packet to the network. This fact implies an additional delay which is

not the simple 20 ms but it is $4 \times 20 = 80$ ms (plus the time of a sample – the first one of the 5^{th} segment), which is a notable delay. In order to decrease such delay (while still use the interleaving strategy) a smaller interleaving factor must be used which, for the same oversampling factor, changes the condition point in such a way that even if the problem remains not ill-conditioned, there is some uncertainty since the new system matrix eigenvalue is not exactly predetermined and the range to which it may belong permits that it will be greater than before (see Fig. 4.8). To overcome this problem the oversampling factor must be reduced. There is thus a trade-off in which the decreasing of the interleaving factor implies the decrease of the oversampling factor (and thus the voice quality) in order to preserve the problem condition in a well-known point. In other words, the increase of the voice quality through the signal robustness and signal reconstruction decreases the intelligibility due to the additional delays and *vice versa*. From this discussion a future work to characterise this trade-off in order to maximise the whole communication intelligibility appears as the natural subsequent way.

Optimising the signals reconstruction by using the perceptual MOS metric – Still concerning the chapter 4, in the signal reconstruction experiences that were done by using the linear interpolation algorithms, the residual error between iterations was used as a stop criterion. To measure this residual error, the RMSE metric was used. Despite the correlation between RMSE and perceptual voice quality, RMSE is not the appropriated metric to measure such a quality, as previously discussed. In this way it may occur that in the convergent reconstruction process, iterations are going to be carried out without expressive gain in what concerns perceptual quality enhancement. It figures out that if the MOS metric would be used to measure the residual error, the reconstruction process could become more effective since the unnecessary precision patent in the RMSE would be blurred by using the MOS and so it would be expected to perform fewer iterations to attain a certain quality.

Implementation of the minimum dimension reconstruction algorithm in a real VoIP communication context – All the tests presented in chapter 4 were carried out by simulation and isolated from a whole, real VoIP system. Despite the good results that were achieved, they do not reflect real conditions of an entire voice communications

system. However, based on the results and conclusions presented in chapter 4 and on the previously proposed interleaving factor study, it figures out the implementation of a complete reconstruction system in a real VoIP system as the natural next step to investigate with more accuracy the contribute of the reconstruction techniques in a real voice communications system. To do this, a margin to introduce some additional delay must be identified so that the reconstruction and interpolation delays could be properly accommodated. The implementation of the reconstruction algorithm in a dedicated hardware, such as Digital Signal Processing (DSP) or Field-Programmable Gate Array (FPGA) as well as the pre-calculation of a pool of system matrices, S, seem to be key factors to the success of such implementation.

Optimising packet classification by using the MOS metric – In the implementation of the classification algorithm described in the chapter 6, the selection of the most important packets is based on the distortion each packet contributes to the distortion of the whole message. In the present case the used distortion measure is the RMSE. The use of this measure has been useful since it permits to be applied to small packets (typically 10 ms, 20 ms) as is the case of VoIP packets and also contributes to the best performance of the algorithm. However, as discussed before (in section 3.2) this kind of objective metric does not provide the better perceptual quality evaluation. In face of this, the next step to enhance the classification task accuracy appears as being to use the MOS as the distortion metric. However, the use of this metric requires that the voice packets are at least 500 ms long as required by the PESQ algorithm. If on the one hand it would promise a better metric based on which decisions about best representative packets are taken, on the other hand it would represent a dramatic increase on the computational delay and effort that may become prohibitive the classification process. The solution to overcome this problem emerges as enhancing the algorithm so that it uses the same packet length but instead of minimising distortions it maximises MOS values and beyond to calculate individual packet distortions it calculates MOS of the whole signal and uses them as feedback to the packet selection process. The changes of the algorithm also include the need to iteratively assemble and disassemble the entire signal formed by its packets. This practice makes arise new timing concerns that must be carefully addressed.

Using linear interpolation algorithms to reconstruct the signal after packet prioritisation – In the section 6.4 a combined technique using packet prioritisation and the Papoulis-Gerchberg (PG) algorithm as means to reconstruct the low priority lost packets was proposed. The presented results permitted to see this hybrid technique as promising since it was possible to save 40% of iterations of the PG algorithm and reduce the reconstruction error by a factor of about 10. Results were presented as preliminary since they must be considered as partial. In fact the computational effort of the classification algorithm as well as the zero-order hold interpolation were not considered. As future work it is worth to define and carry out tests that consider open questions as the referred overall computational effort or timing concerns and extend tests to open the range of signals or packet sizes. The use of the MOS instead or complementing the reconstruction accuracy must also be considered.

* * *

References

[1] Theresa M. Flormata Ballesteros, *Speech and Oral communication.* Katha Publisher Co., Inc., 2003. ISBN: 971-574-069-3.

[2] A. Hart Davis, *History: The Definitive Visual Guide : From the Dawn of Civilization to the Present Day.* Dorling Kindersley, 2007. ISBN: 9780756631192.

[3] Guy Cook, *Applied Linguistics.* Series: Oxford introductions to language study; Applied linguistics and language study, Oxford University Press, 2nd ed., Feb 2003. ISBN: 0-19-437598-6.

[4] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall, 2009.

[5] Theo Zourzouvillys and Eric Rescorla, "An Introduction to Standards-Based VoIP: SIP, RTP, and Friends," *IEEE Internet Computing*, vol. 14, pp. 69–73, Mar/Ap 2010.

[6] T. Chen, J. Walrand, and D. Messerschmitt, "Dynamic priority protocols for packet voice," *Selected Areas in Communications, IEEE Journal on*, vol. 7, pp. 632 –643, Jun 1989.

[7] H. Sanneck, N. T. L. Le, M. Haardt, and W. Mohr, "Selective Packet Prioritization for Wireless Voice over IP," in *4th International Symposium on Wireless Personal Multimedia Communication*, 2001.

[8] F. Beritelli, A. Gallotta, and C. Rametta, "A dual streaming approach for speech quality enhancement of VoIP service over 3G networks," in *Digital Signal Processing (DSP), 2013 18th International Conference on*, pp. 1–5, IEEE, July 2013.

[9] Mousa Al Akhras and Iman Al Momani, *VoIP Technologies*, ch. VoIP Quality Assessment Technologies, pp. 1–36. Intech, Feb 2011. ISBN: 978-953-307-549-5.

[10] Isabel Borges, "VoIP Interconnection Challenges. Workshop do Plano Inovação 2008-2010," Portugal Telecom Inovação, S. A., Aveiro, Mar 2009.

[11] Autorité de Régulation des Communications Électroniques et des Postes, "Observatoire trimestriel des marchés des communications électroniques en France; 3ème trimestre 2010; résultats définitifs." Website, Jan 2011. http://www.arcep.fr. Acceded on Mar 2011.

[12] EntirelyVoIP - Everything VoIP Related, "EntirelyVoIP." Website, April 2008. Acceded on Sep 2009. http://entirelyvoip.com.

[13] www.3G.co.uk, "Cellular VoIP will generate more revenue than all fixed VoIP services." Website, Sep 2006. Acceded on Apr 2009. http://www.3g.co.uk/.

[14] A. Mondal, C. Huang, J. Li, M. Jain, and A. Kuzmanovic, "A Case for WiFi Relay: Improving VoIP Quality for WiFi Users," in *Communications (ICC), 2010 IEEE International Conference on*, pp. 1–5, May 2010.

[15] T. Chakraborty, A. Mukhopadhyay, S. Bhunia, I. Misra, and S. Sanyal, "Analysis and enhancement of QoS in cognitive radio network for efficient VoIP performance," in *Information and Communication Technologies (WICT), 2011 World Congress on*, pp. 904–909, 2011.

[16] TransNexus co., "Four VoIP trends to watch for in 2013." Website, Jan 2013. Acceded on Apr 2013. http://www.transnexus.com/index.php/issue-5-january-2013/four-voip-trends-to-watch-for-in-2013.

[17] Jeff Hecht, "All Smart, No Phone," *IEEE Spectrum*, pp. 30–35, Oct. 2014.

[18] Andrew Odlyzko, "Internet pricing and the history of communications," *Computer Networks*, vol. 36, pp. 493–517, Aug 2001.

[19] Fierce Wireless Europe, "Report: Mobile VoIP users to reach 1 billion by 2017." Website, Jan 2013. Acceded on Apr 2013. http://www.fiercewireless.com/europe/story/report-mobile-voip-users-reach-1-billion-2017/2013-01-02.

[20] C.-H. K. Chu, H. Pant, S. H. Richman, and P. Wu, "Enterprise VoIP Reliability," in *Networks 2006, 12th International Telecommunications Network Strategy and Planning Symposium*, Nov 2006.

[21] M. S. Stephen M. Sacker and Catherine Spence, "The Business Case for Enterprise VoIP," *White Paper, Intel Information Technology. Computer Manufacturing VoIP/SoIP*, Feb. 2006.

[22] ITU-T, "G.114: One-way transmission time," *G series: Transmission systems and media, digital systems and networks*, May. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2003.

[23] G. Anand, R. Vaidya, and T. Velmurugan, "Performance Analysis of VoIP Traffic using various Protocols and Throughput enhancement in WLANs," in *Computer, Communication and Electrical Technology (ICCCET), 2011 International Conference on*, pp. 176–180, 2011.

[24] J. Holub and O. Slavata, "Impact of IP channel parameters on the final quality of the transferred voice," in *Wireless Telecommunications Symposium (WTS), 2012*, pp. 1–5, 2012.

[25] M. Guéguin, R. Le Bouquin Jeannès, V. Gautier Turbin, G. Faucon, and V. Barriac, "On the Evaluation of the Conversational Speech Quality in Telecommunications," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. Article ID 185248, 2008.

[26] M. Voznak, A. Kovac, and M. Halas, "Effective packet loss estimation on VoIP jitter buffer," in *Networking 2012 Workshops*, pp. 157–162, Springer, 2012.

[27] Z. Becvar, J. Zelenka, M. Brada, and L. Novak, "Comparison of Common PLC Methods Used in VoIP Networks," in *Systems, Signals and Image Processing, 2007*, pp. 389 – 392, IEEE, June 2007.

[28] V. Grandcharov and W. B. Kleijn, *Springer Handbook of Speech Processing*, ch. 5: Speech quality assessment, pp. 83–99. Oct 2007. ISBN: 978-3-540-49125-5.

[29] M. Fiedler, T. Hossfeld, and P. Tran Gia, "A generic quantitative relationship between quality of experience and quality of service," *Network, IEEE*, vol. 24, no. 2, pp. 36–41, 2010.

[30] B. Cheetham and K. M. Nasr, "Error concealment for voice over WLAN in converged enterprise networks," in *Proceedings of the IST Mobile Summit*, pp. 307–313, 2006.

[31] D. Goodman, G. Lockhart, O. Wasem, and Wai-Choong Wong, "Waveform substitution techniques for recovering missing speech segments in packet voice communications," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 34, pp. 1440–1448, Dec. 1986.

[32] V. N. Parikh, J.-H. Chen, and G. Aguilar, "Frame erasure concealment using sinusoidal analysis-synthesis and its application to MDCT-based codecs," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 2, pp. II905–II908, IEEE, 2000.

[33] ITU-T, "G.711-Appendix I: A High Quality Low-Complexity Algorithm for Packet Loss Concealment With G.711," *G series: Transmission systems and media, digital*

systems and networks, Sep 1999. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2000.

[34] S. R. Miralavi, S. Ghorshi, M. Mortazavi, and J. Choupan, "Packet loss replacement in VoIP using a recursive low-order autoregressive model-based speech," in *Systems, Signals and Devices (SSD), 2011 8th International Multi-Conference on*, pp. 1–4, IEEE, Mar 2011.

[35] Fatiha Merazka, "Improved Packet Loss Recovery using Interleaving for CELP-type Speech Coders in Packet Networks," *IAENG International Journal of Computer Science*, vol. 36, Feb. 2009.

[36] N. Aoki, "VoIP packet loss concealment based on two-side pitch waveform replication technique using steganography," in *Signal Processing, IEEE Transactions on*, vol. 3, pp. 52– 55, TENCON 2004. 2004 IEEE Region 10 Conference, Nov 2004.

[37] F. Neves, S. Soares, and P. Assunção, "O algoritmo de Papoulis Gerchberg e a reconstrução de voz comutada em tempo real," in *Engenharias '07 - Inovação e desenvolvimento - actas das apresentações*, vol. II, (Covilhã - Portugal), pp. 496–501, Confeng'2007, Universidade da Beira Interior, Nov 2007.

[38] F. Neves, S. Soares, M. Reis, F. Tavares, and P. Assunção, "VoIP reconstruction under a minimum interpolation algorithm," *Consumer Electronics, 2008. ISCE 2008. IEEE International Symposium on*, pp. 1–3, April 2008.

[39] Filipe Neves, Salviano Soares, Pedro Assunção, and Filipe Tavares, *VoIP Technologies*, ch. 3: "Enhanced VoIP by Signal Reconstruction and Voice Quality Assessment", pp. 55–88. Intech, Feb 2011. ISBN: 978-953-307-549-5.

[40] M. I. T. da Costa, F. Neves, S. Soares, and J. Barroso, "Signal Processing Interpolation and Problem Conditioning Educational Workbench," *Computer Applications in Engineering Education*, vol. 7, no. 12, p. p56, 2013.

[41] F. Neves, S. Soares, P. Assunção, F. Tavares, and S. Cardeal, *Sociotechnical Enterprise Information Systems Design and Integration*, ch. 10: Methods for quality assessment in enterprise VoIP communications, pp. 154–170. Business Science Reference (an imprint of IGI Global), 2013. DOI: 10.4018/978-1-4666-3664-4.ch010.

[42] F. Neves, S. Soares, P. Assunção, P., F. Tavares, and S. Cardeal, "Quality Evaluation Methods to Improve Enterprise VoIP Communications," in *ENTERprise Information Systems* (M. M. Cruz-Cunha, J. Varajão, P. Powell, and R. Martinho, eds.), vol. 220 of *Communications in Computer and Information Science*, pp. 111–119, Springer-Verlag Berlin Heidelberg, Oct. 2011. DOI: 10.1007/978-3-642-24355-4_12.

[43] F. Neves, S. Cardeal, S. Soares, P. Assunção, and F. Tavares, "Quality model for monitoring QoE in VoIP services," in *EUROCON-International Conference on Computer as a Tool (EUROCON), 2011 IEEE*, pp. 1–4, IEEE, 2011.

[44] S. Cardeal, F. Neves, S. Soares, F. Tavares, and P. Assunção, "ArQoS®: System to monitor QoS/QoE in VoIP," pp. 1–2, 2011.

[45] Filipe Neves, Salviano Soares, and Pedro Assunção, "Enhanced MOS in VoIP over priority erasure channels based on optimal packet classification," *9th Conference on Telecommunications*, pp. 341–344, May 2013.

[46] ITU-T, "P.862: Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *P series: Terminals and subjective and objective assessment methods*, Feb 2001. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2001.

[47] ITU-T, "G.107: The E-Model, a computational model for use in transmission planning," *G series: Transmission systems and media, digital systems and networks*, Mar 2005. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2005.

[48] J. Janssen, D. De Vleeschauwer, M. Buchli, and G. Petit, "Assessing voice quality in packet-based telephony," *Internet Computing, IEEE*, vol. 6, pp. 48–56, May-Jun 2002.

[49] A. Gabay, M. Kieffer, and P. Duhamel, "Joint Source-Channel Coding Using Real BCH Codes for Robust Image Transmission," *Image Processing, IEEE Transactions on*, vol. 16, pp. 1568–1583, June 2007.

[50] L. Ding and R. Goubran, "Assessment of effects of packet loss on speech quality in VoIP," in *Haptic, Audio and Visual Environments and Their Applications, 2003. HAVE 2003. The 2nd IEEE Internatioal Workshop on*, pp. 49–54, Sep 2003.

[51] Li Mojia, Muqing Wu, Wu Dapeng, Wang Lizhong, and Xu Chunxiu, "Packet Loss Concealment Using Enhanced Waveform Similarity OverLap-and-Add Technique with Management of Gains," in *Wireless Communications, Networking and Mobile Computing, 2009. WiCom '09. 5th International Conference on*, pp. 1–4, 2009.

[52] Emmanuel J. Candès, Justin Romberg, and Terence Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, pp. 489–509, February 2006.

[53] S. Kuroiwa, S. Tsuge, and F. Ren, "A lost speech reconstruction method using linguistic information," *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on.*, pp. 126–130, 30 Oct-1 Nov 2005.

[54] N. Park, H. Kim, M. Jung, S. Lee, and S. Choi, "Burst Packet Loss Concealment Using Multiple Codebooks and Comfort Noise for CELP-Type Speech Coders in Wireless Sensor Networks," *Sensors' 2011*, vol. 11, no. 5, pp. 5323–5336, 2011.

[55] J. Tang and F. Itakura, "Double sided periodic substituion (DSPS) method for recovering missing speech," *ISSPA*, pp. 544–549, 1987.

[56] S. Miura, H. Nakajima, S. Miyabe, S. Makino, T. Yamada, and K. Nakadai, "Restoration of clipped audio signal using recursive vector projection," in *TENCON 2011-2011 IEEE Region 10 Conference*, pp. 394–397, IEEE, 2011.

[57] S.-G. Bae, H.-W. Park, and M.-J. Bae, "A Study on Enhancement of Speech using Non-uniform Sampling," *IJHIT*, vol. 5, no. 2, pp. 237–242, 2012.

[58] N. Erdol, C. Castelluccia, and A. Zilouchian, "Recovery of missing speech packets using the short-time energy and zero-crossing measurements," *IEEE Transactions on Speech and Audio Processing*, vol. 1, pp. 295–303, Jul. 1993.

[59] N. Aoki, "A packet loss concealment technique for VoIP using steganography based on pitch waveform replication," *IEICE Transactions on Communications*, vol. 86, no. 12, pp. 2551–2560, 2003.

[60] P. Wolfe and S. Godsill, "Interpolation of missing data values for audio signal restoration using a Gabor regression model," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 5, pp. v/517–v/520 Vol. 5, 2005.

[61] N. Jayant and S. Christensen, "Effects of Packet Losses in Waveform Coded Speech and Improvements Due to an Odd-Even Sample-Interpolation Procedure," *Communications, IEEE Transactions on [legacy, pre - 1988]*, vol. 29, pp. 101–109, Feb. 1981.

[62] B. Wah, Xiao Su, and Dong Lin, "A survey of error-concealment schemes for real-time audio and video transmissions over the Internet," in *International Symposium on Multimedia Software Engineering*, pp. 17–24, 2000.

[63] P. J. S. Ferreira, "Interpolation and the discrete Papoulis-Gerchberg algorithm," *Signal Processing, IEEE Transactions on*, vol. 42, pp. 2596–2606, Oct 1994.

[64] Levent Tosun and Peter Kabal, "Dynamically Adding Redundancy for Improved

Error Concealment in Packet Voice Coding," in *2005 European Signal Processing Conference - EUSIPCO'2005*, 2005.

[65] Xinwen Mu, Hexin Chen, and Yan Zhao, "A frame erasure concealment method based on pitch and gain linear prediction for AMR-WB codec," in *Consumer Electronics (ICCE), 2011 IEEE International Conference on*, pp. 815–816, 2011.

[66] ITU-T, "G.729: Coding of speech at 8 kbit/s using conjugate structure algebraic-code-excited linear prediction (CS-ACELP)," *G series: Transmission systems and media, digital systems and networks*, Jan 2007. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2008.

[67] ITU-T, "G.729.1: G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729," *G series: Transmission systems and media, digital systems and networks*, May 2006. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2007.

[68] ITU-T, "G.723.1: Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s," *G series: Transmission systems and media, digital systems and networks*, May 2006. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2007.

[69] V. Bhute and U. N. Shrawankar, "Error concealment schemes for speech packet transmission over IP network," in *Systems, Signals and Image Processing, 2008. IWSSIP 2008. 15th International Conference on*, pp. 185–188, IEEE, 2008.

[70] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2, pp. 554–557, Apr 1993.

[71] Y. J. Liang, N. Farber, and B. Girod, "Adaptive playout scheduling and loss concealment for voice communication over IP networks," *Multimedia, IEEE Transactions on*, vol. 5, pp. 532–543, Dec 2003.

[72] S. Grofit and Y. Lavner, "Time-Scale Modification of Audio Signals Using Enhanced WSOLA With Management of Transients," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 106–115, 2008.

[73] Wang Lizhong, Muqing Wu, Li Mojia, and Wei Lulu, "A packet loss concealment method base on GWSOLA algorithm and signification transient detectd," in *Network Infrastructure and Digital Content, 2009. IC-NIDC 2009. IEEE International Conference on*, pp. 745–749, 2009.

[74] J. F. Yeh, M. D. Kuo, and Z. H. Hsu, "Packet Loss Concealment Using Dual-Side Waveform Similarity Overlap-and-Add," *Applied Mechanics and Materials*, vol. 284, pp. 2867–2871, Jan 2013.

[75] Fang Liu, JongWon Kim, and C.-C. Kuo, "Adaptive delay concealment for Internet voice applications with packet based time-scale modification," *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, vol. 3, pp. 1461–1464 vol.3, 2001.

[76] S. B. Moon, J. Kurose, and D. Towsley, "Packet audio playout delay adjustment: performance bounds and algorithms," *Multimedia systems*, vol. 6, no. 1, pp. 17–28, 1998.

[77] B. Balavenkatesh, K. Krishnan, S. Ramkumar, V. Hency, and D. Sridharan, "Enhancement of QoS of VoIP over Heterogeneous Networks by Improving Handoff Speed and Throughput," in *Advances in Computing, Control, Telecommunication Technologies, 2009. ACT '09. International Conference on*, pp. 840–844, 2009.

[78] T. V. Johnson and A. Zhang, "Dynamic playout scheduling algorithms for continuous multimedia streams," *Multimedia Systems*, vol. 7, no. 4, pp. 312–325, 1999.

[79] Y. Liang, N. Farber, and B. Girod, "Adaptive playout scheduling using time-scale modification in packet voice communications," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, vol. 3, pp. 1445–1448 vol.3, 2001.

[80] M. Narbutt, A. Kelly, P. Perry, and L. Murphy, "Adaptive VoIP playout scheduling: assessing user satisfaction," *IEEE Internet Computing*, vol. 9, pp. 28– 34, Jul. 2005.

[81] B. Sat and B. W. Wah, "Playout scheduling and loss-concealments in VoIP for optimizing conversational voice communication quality," in *Proceedings of the 15th international conference on Multimedia*, pp. 137–146, ACM, 2007.

[82] B. Kim, H.-G. Kim, J. Jeong, and J. Kim, "VoIP receiver-based adaptive playout scheduling and packet loss concealment technique," *Consumer Electronics, IEEE Transactions on*, vol. 59, no. 1, pp. 250–258, 2013.

[83] T. Chakraborty, I. Saha Misra, and S. K. Sanyal, "Proactive QoS Enhancement Technique for Efficient VoIP Performance over Wireless LAN and Cognitive Radio Network," *Journal of Networks*, vol. 7, no. 12, pp. 1925–1942, 2012.

[84] P. Jawahar, V. Vaidehi, and D. E. Nirmala, "QoS enhancement in wireless VoIP networks using interactive multiple model based Kalman filter," *Wireless Personal Communications*, vol. 65, no. 1, pp. 67–81, 2012.

[85] Jianying Li, Binyang Xu, Zhangjing Xu, Shaoqian Li, and Yi Liu, "Adaptive Packet Scheduling Algorithm for Cognitive Radio System," in *Communication Technology, 2006. ICCT '06. International Conference on*, pp. 1–5, 2006.

[86] D. Johnson, Y. Hu, and D. Maltz, "RFC 4728 - The Dynamic Source Routing Protocol (DSR) for Mobile Ad-Hoc Networks for IPv4," *[Online]. Available: http://www.rfc-editor.org/rfc/rfc4728.txt*, Feb 2007. Accedded on 2013, Dec.

[87] T. Imielinski and J. Navas, "RFC 2009 - GPS-Based Addressing and Routing,"

[Online]. Available: http://www.rfc-editor.org/rfc/rfc2009.txt, Nov 1996. Acceded on Dec 2013.

[88] C. Perkins, E. Belding Royer, and S. Das, "RFC 3561 - Ad hoc On-Demand Distance Vector (AODV) Routing," *[Online]. Available: http://www.rfc-editor.org/rfc/rfc3561.txt*, Jul 2003. Acceded on Dec 2013.

[89] T. Clausen and P. Jacquet, "RFC 3626 - Optimized Link State Routing Protocol (OLSR)," *[Online]. Available: http://www.rfc-editor.org/rfc/rfc3626.txt*, Oct 2003. Acceded on Dec 2013.

[90] E. Alvarez Flores, J. Ramos Munoz, J. Navarro Ortiz, P. Ameigeiras, and J. Lopez Soler, "User-Level Quality Assessment of a Delay-Aware Packet Dropping Scheme for VoIP," *Network Protocols and Algorithms*, vol. 3, no. 3, pp. 38–66, 2011.

[91] S. Al Rubaye, A. Al Dulaimi, and J. Cosmas, "Cognitive femtocell," *Vehicular Technology Magazine, IEEE*, vol. 6, no. 1, pp. 44–51, 2011.

[92] Stylianos Dimitriou and Vassilis Tsaoussidis, "Promoting effective service differentiation with Size-oriented Queue Management," *Computer Networks*, vol. 54, no. 18, pp. 3360–3372, 2010.

[93] L. Mamatas and V. Tsaoussidis, "Less Impact Better Service (LIBS)," *Annals of Telecommunications*, vol. 65, no. 7, pp. 447–459, 2010.

[94] S. Dimitriou, A. Tsioliaridou, and V. Tsaoussidis, "Introducing size-oriented dropping policies as QoS-supportive functions," *Network and Service Management, IEEE Transactions on*, vol. 7, pp. 14–27, Mar 2010.

[95] Jae-Yul Yoon and Hochong Park, "Improving the Speech Quality of VoIP by Packet Prioritization," *Signal Processing Letters, IEEE*, vol. 18, pp. 725–728, Dec 2011.

[96] Mingyu Chen and M. Murthi, "Optimized unequal error protection for voice over IP," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, vol. 5, pp. V – 865–8, May 2004.

[97] C. Hoene, I. Carreras, and A. Wolisz, "Voice over IP: Improving the quality over wireless LAN by adopting a booster mechanism – an experimental approach," in *ITCom 2001: International Symposium on the Convergence of IT and Communications*, International Society for Optics and Photonics, 2001.

[98] C. Padhye, K. Christensen, and W. Moreno, "A new adaptive FEC loss control algorithm for voice over IP applications," in *Performance, Computing, and Communications Conference, 2000. IPCCC '00. Conference Proceeding of the IEEE International*, pp. 307–313, 2000.

[99] A. Gomez, J. Carmona, A. Peinado, and V. Sanchez, "A Multipulse-Based Forward Error Correction Technique for Robust CELP-Coded Speech Transmission Over Erasure Channels," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1258–1268, 2010.

[100] D. Florencio and Li-Wei He, "Enhanced adaptive playout scheduling and loss concealment techniques for voice over IP networks," in *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on*, pp. 129–132, 2011.

[101] F. Beritelli, A. Gallotta, S. Palazzo, and C. Rametta, "Dual stream transmission to improve mobile VoIP services over HSPA: A practical test bed," in *Image and Signal Processing and Analysis (ISPA), 2013 8th International Symposium on*, pp. 773–777, Sept 2013.

[102] D. Pisoni and S. Hunnicutt, "Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '80.*, vol. 5, pp. 572–575, 1980.

[103] C. V. Pavlovic, M. Rossi, and R. Espesser, "Use of the magnitude estimation technique for assessing the performance of text-to-speech synthesis systems," *The Journal of the Acoustical Society of America*, vol. 87, p. 373, 1990.

[104] E. S. Marta, A. Lopes, F. Neves, Artur Silva, and Carlos Edgar Lopes, "AudiSpeech–A speech audiometry system for assessment and rehabilitation of severaly to profoundy hearing impaired patients," in *Proceedings of the 3rd International Cochlear Implant Conference. Innsbruck, Austria* (E. H. I. J., Hochmair, ed.), Apr 1993.

[105] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual evaluation of voice quality: review, tutorial, and a framework for future research," *Journal of Speech, Language and Hearing Research*, vol. 36, no. 1, p. 21, 1993.

[106] H. Herzel, D. Berry, I. R. Titze, and M. Saleh, "Analysis of vocal disorders with methods from nonlinear dynamics," *Journal of Speech, Language and Hearing Research*, vol. 37, no. 5, p. 1008, 1994.

[107] W. Daumer, "Subjective evaluation of several efficient speech coders," *Communications, IEEE Transactions on*, vol. 30, no. 4, pp. 655–662, 1982.

[108] V. Viswanathan, W. Russell, and A. Huggins, "Objective speech quality evaluation of mediumband and narrowband real-time speech coders," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83*, vol. 8, pp. 543–546, 1983.

[109] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," *Selected Areas in Communications, IEEE Journal on*, vol. 6, no. 2, pp. 242–248, 1988.

[110] A. Mackie, S. Aidarous, S. Mahmoud, and J. Riordon, "Design and performance evaluation of a packet voice system," *Vehicular Technology, IEEE Transactions on*, vol. 32, no. 2, pp. 158–168, 1983.

[111] A. W. Rix and M. P. Hollier, "The perceptual analysis measurement system for robust end-to-end speech quality assessment," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3, pp. 1515–1518, IEEE, 2000.

[112] A. E. Conway, "A passive method for monitoring voice-over-IP call quality with ITU-T objective speech quality measurement methods," in *Communications, 2002. ICC 2002. IEEE International Conference on*, vol. 4, pp. 2583–2586, IEEE, 2002.

[113] S. Mohamed, G. Rubino, and M. Varela, "A method for quantitative evaluation of audio quality over packet networks and its comparison with existing techniques," *Measurement of speech and audio quality in networks (MESAQIN)*, no. 2, p. 0, 2004.

[114] S. Rein, F. H. Fitzek, and M. Reisslein, "Voice quality evaluation in wireless packet communication systems: a tutorial and performance results for RHC," *Wireless Communications, IEEE*, vol. 12, no. 1, pp. 60–67, 2005.

[115] R. Guski, "Psychological methods for evaluating sound quality and assessing acoustic information," *Acta Acustica united with Acustica*, vol. 83, no. 5, pp. 765–774, 1997.

[116] W. Voiers, "Evaluating processed speech using the diagnostic rhyme test," *Speech Technology*, vol. 1, no. 4, pp. 30–39, 1983.

[117] P. Combescure, A. Le Guyader, and A. Gilloire, "Quality evaluation of 32 kbit/s coded speech by means of degradation category ratings," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82*, vol. 7, pp. 988–991, IEEE, 1982.

[118] ITU-T, "P.800: Methods for subjective determination of transmission quality," *P series: Terminals and subjective and objective assessment methods*, Aug 1996. Telecommunication Standardization Sector of International Telecommunication Union. Amended at Helsinki, 1993; revised in Geneva, 1996.

[119] J. G. Beerends and J. A. Stemerdink, "A perceptual speech-quality measure based on a psychoacoustic sound representation," *Journal of the Audio Engineering Society*, vol. 42, no. 3, pp. 115–123, 1994.

[120] M. P. Hollier, M. Hawksford, and D. Guard, "Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain," in *IEE Proceedings-Vision, Image and Signal Processing*, vol. 141, pp. 203–208, IET, 1994.

[121] A. Rix, R. Reynolds, and M. Hollier, "Perceptual Measurement of End-to-End Speech Quality Over Audio and Packet-Based Networks," in *Audio Engineering Society Convention 106*, 5 1999.

[122] J. G. Beerends, B. Busz, P. Oudshoorn, J. Van Vugt, K. Ahmed, and O. Niamut, "Degradation Decomposition of the Perceived Quality of Speech Signals on the Basis of a Perceptual Modeling Approach," *J. Audio Eng. Soc*, vol. 55, no. 12, pp. 1059–1076, 2007.

[123] M. Barkowsky, J. Bialkowski, R. Bitto, and A. Kaup, "Temporal registration using 3D phase correlation and a maximum likelihood approach in the perceptual evaluation of video quality," *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, pp. 195–198, 2007.

[124] ITU-T, "P.863: Perceptual objective listening quality assessment," *P series: Terminals and subjective and objective assessment methods*, Jan 2011. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2011.

[125] T. H. Falk and W.-Y. Chan, "Performance study of objective speech quality measurement for modern wireless-VoIP communications," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, no. Article ID 104382, p. 12, 2009.

[126] John G. Beerends, Andries P. Hekstra, Antony W. Rix, and Michael P. Hollier, "Perceptual Evaluation of Speech Quality (PESQ) – The New ITU Standard for End-to-End Speech Quality Assessment - Part II - Psychoacoustic Model," *AES journal of the Audio Engineering Society Audio / Acoustics / Applications*, vol. 50, pp. 765–778, Oct 2002.

[127] Jinhe Zhou, Tonghai Wu, and Junmin Leng, "Research on voice codec algorithms of SIP phone based on embedded system," in *Wireless Communications, Networking and Information Security (WCNIS), 2010 IEEE International Conference on*, pp. 183–187, 2010.

[128] T. Hoßfeld and A. Binzenhöfer, "Analysis of Skype VoIP traffic in UMTS: End-to-end QoS and QoE measurements," *Computer Networks*, vol. 52, no. 3, pp. 650–666, 2008.

[129] ITU-T, "P.800.1: Mean Opinion Score (MOS) terminology," *P series: Terminals and subjective and objective assessment methods*, Jul 2006. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2006.

[130] ITU-T, "P.880: Continuous evaluation of time varying speech quality," *P series: Terminals and subjective and objective assessment methods*, May 2004. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2004.

[131] A. Rix, J. Beerends, Doh-Suk Kim, P. Kroon, and O. Ghitza, "Objective Assessment of Speech and Audio Quality - Technology and Applications," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 1890–1901, Nov 2006.

[132] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *Selected Areas in Communications, IEEE Journal on*, vol. 10, no. 5, pp. 819–829, 1992.

[133] J. Beerends and J. Stemerdink, "A perceptual audio quality measure based on a psychoacoustic sound representation," *Journal of the Audio Engineering Society*, vol. 40, pp. 963–978, Dec 1992.

[134] ITU-T, "P.861: Objective quality measurement of telephoneband (300-3400 Hz) speech codecs," *P series: Terminals and subjective and objective assessment methods*, Feb 1998.

[135] A. Rix and M. Hollier, "Robust design methodology for telephony assessment models," *ITU-T COM*, pp. 12–D031, 1998.

[136] P. Juric, "An Objective Speech Quality Measurement in the QVoice," in *Proceedings of the IEEE 5th International Workshop on Systems, Signals and Image Processing (IWSSIP)*, pp. 156–163, 1998.

[137] A. Anskaitis and A. Kajackas, "The Tool for Quality Estimation of Short Voice Segments," *Electronics and Electrical Engineering. –Kaunas: Technologija*, no. 8, p. 104, 2010.

[138] ITU-T, "P.862.1: Mapping function for transforming P.862 raw result scores to MOS-LQO," *P series: Terminals and subjective and objective assessment methods*, Nov 2003. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2004.

[139] ITU-T, "P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," *P series: Terminals and subjective and objective assessment methods*, Nov 2007. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2008.

[140] ITU-T, "P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications," *P series: Terminals and subjective and objective assessment methods*, May 2004. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2005.

[141] ITU-T, "G.108: Application of the E-model: A planning guide," *G series: Transmission systems and media, digital systems and networks*, Sep 1999. Telecommunication Standardization Sector of International Telecommunication Union. Printed in Geneva, 2000.

[142] ITU-T, "G.109: Definition of categories of speech transmission quality," *G series: Transmission systems and media, digital systems and networks*, Sep 1999.

Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 1999.

[143] Pierre-Gérard Fontolliet, *Systèmes de Télécommunications*, ch. 2: Planification: objectifs, constraintes, méthodes, pp. 23–65. Traité d'Électricité - Vol XVIII, Presses Polytechniques et Universitaires Romandes, 2ème ed., 1994. ISBN: 2-88074-269-2.

[144] ITU-T, "Rec. G.113: Transmission impairments due to speech processing," *G series: Transmission systems and media, digital systems and networks*, Nov 2007. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2007.

[145] David Kincaid and Ward Cheney, *Numerical Analysis: Mathematics of Scientific Computing*, ch. 4: Solving Systems of Linear Equations, pp. 139–253. The Brooks Cole - Thompson Learning, 3rd ed., 2002. ISBN: 0-534-38905-8.

[146] Roger A. Horn and Charles R. Johnson, *Matrix Analysis*. Press Syndicate of the University of Cambridge, 1985. ISBN: 978-0-521-38632-6.

[147] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, *Numerical recipes: the art of scientific computing*. Cambridge University Press, 3rd ed., 2007. ISBN: 978-0-521-88068-8.

[148] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, *Numerical Recipes in C: The art of scientific computation*. Press Sybdicate of Cambridge University Press, 1994. ISBN: 0-521-43108-5.

[149] P. J. S. Ferreira, "Noniterative and fast iterative methods for interpolation and extrapolation," *Signal Processing, IEEE Transactions on*, vol. 42, pp. 3278–3282, Nov 1994.

[150] P. J. S. Ferreira, "The stability of a procedure for the recovery of lost samples in band-limited signals," *Signal Processing, IEEE Transactions on*, vol. 40, pp. 195–205, Dec 1994.

[151] P. Ferreira, "Interpolation in the time and frequency domains," *Signal Processing Letters, IEEE*, vol. 3, pp. 176–178, Jun 1996.

[152] Paulo J.S.G. Ferreira, "Mathematics for Multimedia Signal Processing II: Discrete Finite Frames and Signal Reconstruction," *Signal Processing for Multimedia*, pp. 35–54, 1999.

[153] R. Gerchberg, "Super-resolution through error energy reduction," *Journal of Modern Optics*, vol. 21, no. 9, pp. 709–720, 1974.

[154] A. Papoulis, "A new algorithm in spectral analysis and band-limited extrapolation," *Circuits and Systems, IEEE Transactions on*, vol. 22, pp. 735–742, Sep 1975.

[155] F. Neves, S. Soares, M. C. Reis, F. Tavares, and P. Assunção, "VoIP reconstruction under a minimum interpolation algorithm," in *Consumer Electronics, 2008. ISCE 2008. IEEE International Symposium on*, pp. 1–3, Apr 2008.

[156] ITU-T, "G.111: Loudness Ratings (LRs) in an international connection," *G series: Transmission systems and media, digital systems and networks*, Mar 1993. Telecommunication Standardization Sector of International Telecommunication Union. Helsinki, 1993.

[157] ITU-T, "G.121: Loudness Ratings (LRs) of National Systems," *G series: Transmission systems and media, digital systems and networks*, Mar 1993. Telecommunication Standardization Sector of International Telecommunication Union. Geneva, 1985; amended at Helsinki, 1993.

[158] Pierre-Gérard Fontolliet, *Systèmes de Télécommunications*, ch. 4: Procédés de transmision, pp. 121–150. Traité d'Électricité - Vol XVIII, Presses Polytechniques et Universitaires Romandes, 2ème ed., 1994. ISBN: 2-88074-269-2.

[159] TIA, "TIA TSB-116A: Telecommunications -IP Telephony Equipment - Voice Quality Recommendations for IP Telephony," *TIA Telecommunications Systems Bulletin*, Mar 2006. USA, Arlington, 2006.

[160] ITU-T, "P.564 - Conformance testing for voice over IP transmission quality assessment models," *P series: Terminals and subjective and objective assessment methods*, Nov 2007. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2008.

[161] ITU-T, "P.862.3: Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2," *P series: Terminals and subjective and objective assessment methods*, Nov 2005. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2006.

[162] ITU-T, "P.501: Test signals for use in telephonometry," *P series: Terminals and subjective and objective assessment methods*, Jun 2007. Telecommunication Standardization Sector of International Telecommunication Union. Switzerland, Geneva, 2008.

[163] Z. Li, G. Schuster, A. Katsaggelos, and B. Gandhi, "Rate-distortion optimal video summary generation," *Image Processing, IEEE Transactions on*, vol. 14, pp. 1550–1560, Oct 2005.

[164] G. Nemhauser and L. Wolsey, *Integer and combinatorial optimization*, vol. 18. Wiley New York, 1988.

[165] T. Cormen, C. Leiserson, and R. E. M. P. . M.-H. B. C. Rivest, *Introduction to Algorithms*, ch. 17: Greedy Algorithms, pp. 329–355. The MIT Press, Cambrige, Massachutes, 1990. ISBN: 0-262-03141-8.

[166] Guido M. Schuster and Aggelos K. Katsaggelos, *Rate-Distortion Based Video Compression – Optimal Video Frame Compression and Object Boundary Enconding*, ch. 4: Background, pp. 43–72. Kluwer Academic Publishers, 1997. ISBN: 0-7923-9850-5.

[167] T. Cormen, C. Leiserson, and R. E. M. P. . M.-H. B. C. Rivest, *Introduction to Algorithms*, ch. 16: Dynamic Programming, pp. 301–328. The MIT Press, Cambrige, Massachutes, 1990. ISBN: 0-262-03141-8.

[168] Z. Li, G. Schuster, A. Katsaggelos, and B. Gandhi, "Rate-Distortion optimal video summarization: A dynamic programming solution," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 3, pp. iii–457, IEEE, 2004.

[169] G. Schuster, G. Melnikov, and A. Katsaggelos, "A review of the minimum maximum criterion for optimal bit allocation among dependent quantizers," *Multimedia, IEEE Transactions on*, vol. 1, pp. 3–17, Aug 1999.

[170] Zhongbo Li, Shenghui Zhao, Stefan Bruhn, Jing Wang, and Jingming Kuang, "Comparison and optimization of packet loss recovery methods based on AMR-WB for VoIP," *Speech Communication*, vol. 54, no. 8, pp. 957–974, 2012.

[171] H. P. Singh, S. Singh, and J. Singh, "Computer Modeling & Performance Analysis of VoIP under Different Strategic Conditions," in *Computer Engineering and Applications (ICCEA), 2010 Second International Conference on*, vol. 1, pp. 611–615, IEEE, 2010.

[172] G. Haßlinger and O. Hohlfeld, "The Gilbert-Elliott model for packet loss in real time services on the Internet," in *Measuring, Modelling and Evaluation of Computer and Communication Systems (MMB), 2008 14th GI/ITG Conference-*, pp. 1–15, VDE, 2008.

[173] M. Yajnik, S. Moon, J. Kurose, and D. Towsley, "Measurement and modelling of the temporal dependence in packet loss," in *INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 1, pp. 345–352, IEEE, 1999.

[174] W. Jiang and H. Schulzrinne, "Comparison and optimization of packet loss repair methods on VoIP perceived quality under bursty loss," in *Proceedings of the 12th international workshop on Network and operating systems support for digital audio and video*, pp. 73–81, ACM, 2002.

[175] B. Ngamwongwattana and A. Sombun, "Enhancing adaptive control in VoIP by characterizing congestive packet loss," in *Advanced Communication Technology, 2009. ICACT 2009. 11th International Conference on*, vol. 3, pp. 1847–1851, IEEE, Feb 2009.

[176] A. Clark, "Modeling the effects of burst packet loss and recency on subjective voice quality," in *Proc. IP Telephony Workshop*, Apr 2001.

[177] C. Tie and W. Xing, "Adaptive Playout Buffer Algorithm Based on the Speech Quality Prediction for VoIP Applications," in *2011 International Conference in Electrics, Communication and Automatic Control Proceedings* (R. Chen, ed.), pp. 553–559, Springer, Springer New York, 2012.

[178] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied linear statistical models*. Irwin Chicago, 1996. ISBN: 0-256-11736-5.

[179] National Communications System Technology & Standards Division, "Federal Standard 1037C," *Telecommunications: glossary of telecommunication terms*, Jun 1996. Published by General Services Administration Information Technology Service, United States Department of Commerce.

[180] A. Raake, "Short-and long-term packet loss behavior: towards speech quality prediction for arbitrary loss distributions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 1957–1968, Nov 2006.

[181] S. Jelassi and G. Rubino, "A comparison study of automatic speech quality assessors sensitive to packet loss burstiness," in *Consumer Communications and Networking Conference (CCNC), 2011 IEEE*, pp. 415–420, IEEE, 2011.

[182] J.-C. Bolot and A. Vega Garcia, "Control mechanisms for packet audio in the Internet," in *INFOCOM '96. Fifteenth Annual Joint Conference of the IEEE Computer Societies. Networking the Next Generation. Proceedings IEEE.*, vol. 1, pp. 232–239, Mar 1996.

[183] W. Jiang and H. Schulzrinne, "QoS measurement of Internet real-time multimedia services," in *Proc. NOSSDAV*, 1999.

[184] H. Toral, D. Torres, C. Hernández, and L. Estrada, "Self-Similarity, Packet Loss, Jitter, and Packet Size: Empirical Relationships for VoIP," in *Electronics, Communications and Computers, 2008. CONIELECOMP 2008, 18th International Conference on*, pp. 11–16, IEEE, Mar 2008.

[185] S. Salsano, F. Ludovici, and A. Ordine, "Definition of a general and intuitive loss model for packet networks and its implementation in the Netem module in the Linux kernel," *University of Rome "Tor Vergata"*, pp. 1–60, Sep. 2010.

[186] Networking Group, "NetemCLG." Website, 2009. Acceded on Jan 2012. http://netgroup.uniroma2.it/twiki/bin/view.cgi/Main/NetemCLG#SqngenOver.

The derived model algorithm

This appendix presents the modules that constitute the derived model in an algorithmic language as it is defined in the *algoritmicx* package of LaTeX. It is presented in four parts. Part I, labeled as **Algorithm 1**, describes the procedure of launching the ArQoS© and registering the provided values. It is the part that provides the QoS and QoE values to be used in each of the three derived modules. Part II, labeled as **Algorithm 2**, describes the calculations relative to the local calling area module. The Part III, labeled as **Algorithm 3**, describes the calculations relative to the far-end calling area module and the Part IV, labeled as **Algorithm 4** describes the VoIP module by means of the polynomial functions that approximate the modified E-Model MOS to the reference.

Algorithm 1 Part I - Launch ArQoS system and read values

1: ArQoS injects -10 dBm0 signal
2: $Level\_at\_Side\_A \leftarrow$ ArQoS ▷ Level measured at Side A
3: $Level\_at\_Side\_B \leftarrow$ ArQoS ▷ Level measured at Side B
4: $T \leftarrow$ ArQoS
5: $Ta \leftarrow$ ArQoS
6: $Tr \leftarrow$ ArQoS
7: $Noise\_at\_Side\_A \leftarrow$ ArQoS ▷ Noise measured at Side A
8: $Noise\_at\_Side\_B \leftarrow$ ArQoS ▷ Noise measured at Side B
9: $MOS_{LQE} \leftarrow$ ArQoS

A.1 The local calling area case

Algorithm 2 Part II - Case of a local calling area scenario

10: **if** switches are different **then**
11: **switch** *switch model* **do**
12: **case** *Siemens*
13: $SLR \leftarrow (-10 - Level\_at\_Side\_A + Default\ SLR)$
14: $RLR \leftarrow (-10 - Level\_at\_Side\_B\_Side + Default\ RLR)$
15: $WEPL \leftarrow 110$
16: $R \leftarrow E\text{-}Model(SLR, RLR, WEPL, remaining\ default\ values)$
17: **case** *Alcatel*
18: $SLR \leftarrow (-10 - Level\_at\_Side\_A)$
19: $RLR \leftarrow (-10 - Level\_at\_Side\_B)$
20: $WEPL \leftarrow 0$
21: $T \leftarrow Ta \leftarrow Tr \leftarrow 0$
22: $R \leftarrow E\text{-}Model(SLR, RLR, WEPL, T, Ta, Tr, remaining\ default\ values)$
23: **else**
24: $SLR \leftarrow (-10 - Level\_at\_Side\_A)$
25: $RLR \leftarrow (-10 - Level\_at\_Side\_B)$
26: $WEPL \leftarrow 0$
27: $T \leftarrow Ta \leftarrow Tr \leftarrow 0$
28: $R \leftarrow E\text{-}Model(SLR, RLR, WEPL, T, Ta, Tr, remaining\ default\ values)$
29: $MOS \leftarrow Calculate\_MOS(R)$

A.2 The far-end calling case

Algorithm 3 Part III - Case of a far-end calling scenario

30: **if** NOT $wants\ a\ trade\text{-}off$ **then**
31: **if** switches are different **then**
32: **switch** *switch model* **do**
33: **case** Siemens
34: $SLR\_A \leftarrow (-10 - Level\_at\_Side\_A) + DefaultSLR$
35: $RLR\_B \leftarrow (-10 - Level\_at\_Side\_B) + DefaultRLR$
36: $Nc \leftarrow (Noise\_at\_Side\_A - SLR\_A + Noise\_at\_Side\_A - SLR\_B)/2$
37: $WEPL \leftarrow SLR\_A + RLR\_B + SLR\_B + RLR\_A + 17 + 17$
38: $TELR \leftarrow SLR\_A + RLR\_B + SLR\_B + RLR\_A + 17$
39: $Tr \leftarrow 2 \times Ta$; $T \leftarrow Ta \leftarrow measuredArQoSvalue$
40: $R \leftarrow E\text{-}Model(SLR\_A, RLR\_B, WEPL, TELR, T, Ta, Tr, remaining\ default\ values)$
41: **case** Alcatel
42: $SLR\_A \leftarrow -10 - Level\_at\_Side\_A + Default\_SLR$
43: $RLR\_B \leftarrow -10 - Level\_at\_Side\_B + Default\_RLR$
44: $Nc \leftarrow -70$; $WEPL \leftarrow TELR \leftarrow 5$
45: $T \leftarrow Ta \leftarrow Tr \leftarrow 0$
46: $R \leftarrow E\text{-}Model(SLR\_A, RLR\_B, Nc, WEPL, TELR, T, Ta, Tr, remaining\ default\ values)$
47: **else**
48: $SLR\_A \leftarrow -10 - Level\_at\_Side\_A + Default\ SLR$
49: $RLR\_B \leftarrow -10 - Level\_at\_Side\_A + Default\ RLR$
50: $WEPL \leftarrow TELR \leftarrow 5$
51: $R \leftarrow E\text{-}Model(SLR\_A, RLR\_B, WEPL, TELR, remaining\ default\ values)$
52: **else**
53: $SLR \leftarrow -10 - measured\_level\_Sending\_Side$
54: $SLR \leftarrow -10 - measured\_level\_Receiving\_Side$
55: $WEPL \leftarrow T \leftarrow Ta \leftarrow Tr \leftarrow 0$
56: $R \leftarrow E\text{-}Model(SLR, RLR, WEPL, T, Ta, Tr, remaing\ default\ values)$
57: $MOS \leftarrow Calculate\_MOS(R)$

A.3 The VoIP case

Algorithm 4 Part IV - Case of VoIP scenario

58: **switch** *codec* **do**

59: **case** $G.711$

60: $MOS_{LQO} \leftarrow -0.0058MOS^4_{LQE} + 0.1252MOS^3_{LQE} - 0.6467MOS^2_{LQE} + +1.9197MOS_{LQE} - 0.291$

61: **case** $G.729$

62: $MOS_{LQO} \leftarrow -0.0554MOS^5_{LQE} - 0.7496MOS^4_{LQE} + 3.9507MOS^3_{LQE} - -9.874MOS^2_{LQE} + 11.939MOS_{LQE} - 3.8293$

63: **case**

64: $MOS_{LQO} \leftarrow 0.0018MOS^4_{LQE} + 0.0248MOS^3_{LQE} - 0.4262MOS^2_{LQE} + +2.1953MOS_{LQE} - 0.2914$

B The segment classification algorithm

Algorithm 5 Classify Segments

// Phase 1: Identifies candidates and calculates edge costs

1: **for** $t \leftarrow 1,\ m$ **do**
2: **for** $k \leftarrow (t-1),\ (n-1-m+t)$ **do**
3: $D_t^k \leftarrow \min_{l_{t-2}} \left\{ D_{t-1}^{l_{t-2}} - e^{l_{t-2},k} \right\}$
4: $e^{l_{t-2},k} \leftarrow \sum_{j=k}^{n-1} \left[d(p_j, p_{l_{t-2}}) - d(p_j, p_k) \right]$
5: $p_k \leftarrow \arg D_t^k$

// Phase 2: Selects definitive segments.

6: $p_{m-1} = p_k^* \leftarrow \arg\min_k \{D_m^k\}$ // First, the last segment of M^*...
7: **for** $t \leftarrow (m-1),\ 2$ **do** // ... then remaining segments, except first one.
8: **for** $k \leftarrow (t-1),\ (n-1-m+t)$ **do**
9: $p_k \leftarrow \{D_t^k - e^{k,l_t}\}$
10: $p_{t-1} = p_k^* = \arg\min_k \{p_k\}$
11: $p_{l_0} \leftarrow p_0$ // First segment is p_0, by definition.