Master Thesis

DELAY IMPACTS ON HUMAN-TO-HUMAN MEDIATED INTERACTION AND END-USER QUALITY PERCEPTION

Oscar Golderos Blanco

Institut für Nachrichtentechnik und Hochfrequenztechnik
Fakultät für Elektrotechnik und Informationstechnik
Technische Universität Wien
Univ.Prof. Dipl.-Ing. Dr.techn. Markus Rupp
Supervisor: Dipl.-Ing. Dr. techn. Philipp Svoboda
Vienna, April 2010

Kurzfassung

Abstract

This Master Thesis analyzes the human to human mediated communications and how the mediated conversation is affected by the communications defects. It deals with the delay impacts on communications and the end-user quality perception.

An experimental setup is used which allows adjustment of delay network parameter to study their impact on the perceived connection quality. In this work it has been developed several algorithms that permit to obtain relevant statistics related to each scenario and network condition. Thus, it has been built a test bed that permits to obtain subjective (MOS) and objective parameters in order to analyze the delay effects on speech quality.

It has been designed 15 test scenarios in which there are different degrees of interactivity. A practice test has been performed involving 34 test persons. The results of the sessions test are showed and commented using a subjective and objective point of view.

Finally, this work analyzed the results obtained so we can compare the delay impact as the one specified in the ITU recommendations

Acknowledgments

I would like to give thanks to my family for all the support with this engineering and this project.

Also to Sebastian Egger for his supervision and giving his help in all moment with all my problems whatever they are.

Also to all the friends that participated in my tests: Alberto, Aritz, Ana, Luis, Pol, Juan, Ignasi, Fran, David, Jonathan, Javi, Elena, Sofía, Jorge, Paco pepe, Ferreiro, Raquel, Rocío, Alfonso, Adrienn, Anne-Marie, Miguel and Agnes. And also to those that couldn't participe, Marta, Fabian, Mikel, Laura, Heikki, Ivan, Jorge and Juan.

To Kathrin for her help in the exhausting tests.

I will never forget all friends that have accompanied me in these years of study, Álvaro, Juanillo, Javi, Miky, Chema, César, Navas, Miguel, Keko, Alber, Duplá, Gonzalo, Gerard, Joaquin, Elena and Pablo.

To my teacher Luis Hernandez of the Universidad Politécnica de Madrid for his contribution to the synchronization of the recordings and as well to all the teachers that have helped me in my degree.

Vienna, May 2010 Oscar Golderos Blanco

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Originally, packet-switched networks were constructed for data transmission. In this network, the main requirement was a reliable transmission, so no data would be lost. The Transmission Control Protocol (TCP) assures that every packet is received at the destination. If any packet is lost in the network, the source must repeat sending the lost packet until it is finally received. The transport reliability results in severe latency caused by the transport protocol.

In real time applications, like VoIP, cannot be used TCP because the requirements of low latency does not allow packets to be resent. Thus, the speech packets are sent in real time using the User Datagram Protocol (UDP).

A voice over IP network is composed of several parts. These are :

IP backbone : provides connectivity.

Access network : facilitates the connection between the backbone network and the end user.

Gateway : permits the communications between the packet-switched network and the circuit-switched network.

Terminal : each participant in the VoIP service needs a terminal which may either be a VoIP phone or a computer with a VoIP software client.

The number of multimedia applications running over the Internet has been steadily increasing recently. Voice over IP, IP telephony, audio streaming, video-conferencing, are becoming commonplace. Nevertheless, there are still major challenges to be overcome in order to provide the end user with acceptable levels of service quality, especially over connections with long propagation delays, great differences in channel speeds (from high backbone speeds to comparatively low last mile speeds) and physical transmission media. One of the major issues these applications face is to maximize the perceived QoS (Quality of Experience, QoE) for drastically varying network states.

Since quality is not guaranteed in the current Internet, it is important for the QoS to be constantly monitored so that the applications can take the proper actions needed to mantain it over some minimum level. Therefore, it is essential to determine what are the parameters that mostly influence the user's perception of the QoS and to understand their combined effects on quality from the user's perspective. I will explain that the conversational quality , as perceived by the end user, depends on a complex combination of several parameters.

Subjective quality assesment methods measure the perceived quality from the end-user's perspective. For interactive multimedia streams, the [1, ITU-T P.800] and [2, ITU-T P.920] recommendations give guidelines on how subjective assesment should be performed, define the environmental setup and provide information on the kinds of tasks that the test subjects should perform. The results of these kind of tests are mean opinion scores (MOS), which gives a numeric expression of subjective quality. Normally, subjective tests involve a relatively large group of subjects who (in the case of conversational quality assessment) carry out a conversation (normally based on some tasks) over the system to be tested, and then grade the quality as they perceived it.

The ITU recommendation suggest a 5-point scale, which spans from bad to very good quality.

The [3, ITU-T E.800] gives a formal definition of Quality of Service (QoS):

> "The collective effect of service performance which determines the degree of satisfaction of an user of the service".

The service performance includes four combined aspects:

Service support performance: describes the ability of an organization to provide a service and assist in its utilization.

Service operability performance: indicates the ability of a service to be successfully and easily operated by an user.

Serveability performance: indicates the ability of a service to be obtained within specified tolerances and other given conditions when requested by the user and continue to be provided without excessive impairment for a requested duration.

Service security performance: specifies the protection provided against unauthorized monitoring, fraudulent use, malicious impairment, human mistakes and natural disaster.

We can distinguish, from a technical point of view, two types of QoS:

Network QoS: is described by the parameters that determine the level of performance of the underlying network. In case of a VoIP network, these parameters are packet loss rate, packet transmission delay and jitter.

Terminal QoS: based on the speech codec in use, the packet loss concealment algorithm, the playout buffer mechanism, accoustic properties of the terminal and echo cancelation.

The [4, ITU-T SG12] defines "Quality of Experience" (QoE) as:

"a measure of the overall acceptability of an application or service, as perceived subjectively by the end user".

Since the user acceptability is a crucial issue in order to provide a successful VoIP services, the network and service providers need to mantain an acceptable level of QoE. The packet loss and delay are considered as the most important VoIP network QoS parameters.

The goal of this work is to investigate the impact of the delay on the conversational structure and on the speech quality for different conversation scenarios. I will use an experimental environment that allows simulation of delay in a network and several algorithms that will facilitate relevant statistics about the dialogues and different network conditions used in each one of the scenarios.

1.1.1 Problem Description

The transmission of voice over IP networks (VoIP) has a very important role and an increasingly widespread use. Users expect to achieve their expectations for quality of service regardless of the technology used. In this sense, the "Quality of Experience" (QoE) measured as a network service meets the expectations and needs seen by the user. Moreover, the "Quality of Service" (QoS) is concerned with the measurement of network performance from a technical viewpoint and the ability to manage it to meet the performance required for applications.

VoIP faces the typical problems of data networks that manifest as degradation in service quality perceived by users (QoE). These impairments may be due to delays, jitter (delay difference) and packet loss, basically.

The complete delay is made up of a number of individual components in both the IP-network and the terminal in use. The major components defined by [5, Hammer, 2006] are:

Speech coding delay: coding a digitized speech signal into a bitstream takes proccesing time which depends on the coding technique. Most of today's codecs collect a block of samples resulting in a certain basic delay time. The algorithms of several types of codecs [6,] (ITU-T G.729, AMR) adds a delay between 20 ms and 30 ms.

Packetization: The packetization delay represents the time needed to prepare the speech frames for RTP/UDP/IP transport. This delay depends, partly, of the packet lenght.

Serialization: serialization delay is the fixed amount of time needed to transmit packet frames of a certain size over a link at a certain bandwith.

ADSL transmission/processing delays: ADSL provides a "fast path" or a "slow path" for data transmission. In the slow path, an interleaver is used to improve the protection against burst noise on the DSL link. The delay produced by interleaving depends on interleave depth.

Radio link delay: the GSM radio link introduces 95 ms one-way delay from the acoustic reference point to the PSTN point of inter connect. Thus, deducting the coding delay of 40 ms from the total radio link delay of 40 ms from

the total radio link delay, the channel coding delay and serialization delay of a radio link is about 55 ms.

Propagation delay (backbone): the mean one-way delay for optical fibre systems is aprox 5 microseconds/km. For copper the propagation delay remains low for low/middle distance calls. For example, over a distance of 600 km results in 3 ms of propagation delay.

Queueing delay: in routers and gateways, voice frames are queued for transmission. Due to the variable states of the queues, the queueing delay is variable and contributes esentially to delay jitter.

VoIP gateway delay: VoIP gateways connect IP networks with other networks like PSTN or GSM. Due to the use of different voice coding algorithms, the speech information has to be converted into an appropiate format. The transcoding not only results in additional delay, but also degrades the speech quality.

User Terminal: assuming a PC as the user terminal, an essential amount of latency is introduced by the computer equipment and software. This amount includes the playout buffering, sound card latency, operating system latency and the potential delay of the sound wave from the loudspeaker to the ear of the user.

Additionaly, we have to consider the Call setup delay. It represents the time the user has to wait for a connection to be established after dialing a phone number and influences the user's communication experience. In a typical example, the descomposition of the end to end delay into its components is:

Delay type	Delay value (ms)
Coding	45
Packetization	20
Serialization	10
Propagation	5
Queueing	10
Playout buffer	60
User terminal	30
Total	180

Table 1.1: Types of delays

The International Telecommunication Union (ITU) indicates in their recommendation [7, G.114] the issues of one-way transmission time. This recommendation concern the need to consider the delay impact in today's telecommunications applications and the avoidance of delay whenever possible. [7, G.114] recommends three areas of limits for one way transmission delay:

Delay range (ms)	Description
0-150	Acceptable for most user applications
150 - 400	Acceptable provided that administrations are aware of the transmission time impact on the transmission quality of user applications
Above 400	Unacceptable for general network planning purposes. However it is recognized that in some exceptions cases this limit will be exceeded.

Table 1.2: Delay ranges

The first area basically does not influence a telephone conversation, except for highly interactive tasks. Until 400 ms delay, transmission quality can be accepted for international connection with satellite hops and one-way delays beyond 400 ms are generally unacceptable. The delay impact indicated in ITU-T G.114 are related with results obtained in severals experiments like [8, Kitawaki, 1991].

In this thesis, I investigate about the delay impact on the speech quality. I will compare my results with the [7, ITU-T G.114] recommendation states.

For VoIP, technology can be used in bulk and trade is essential to ensure acceptable voice quality. For this have been developed variable measure methods. These methods are divided into subjective and objective ones. The subjective methods of measuring the quality of service are based on direct knowledge of the feedback from users. Typically results in an average of views (MOS - Mean Opinion Score). The objective methods measure physical properties of a network to predict or estimate the performance perceived by users. In turn they are divided into intrusive and nonintrusive. Intrusive methods involve the injection of a known speech signal in the channel and the study of degradation at the output, for example PESQ. Nonintrusive methods monitor certain parameters at a point in the network, and based on these, they allow to establish the quality perceived by the user in real time.

1.1.2 Goals of this work

This Master Thesis will analyze the human to human mediated communication and how the mediated conversation is affected by the communication defects. It will deal with the delay impacts on communications and the end-user quality perception.

I will use an experimental environment where we can control network parameters like delay. The different network conditions will be applied for each scenario in order to analyze the impact on the correspondent interaction. I will develop several algorithms that will permit to obtain relevant statistics related to each scenario and each network conditions. These statistics are:

1. Speaker alternation rate
2. Double talk
3. Mutual silence

According to the ITU recommendations [9, P800.1], and [2, P920], I designed different types of scenarios in which there are several degrees of interactivity. A user test was performed to a certain number of participants using the scenarios described. These tests were performed in the i:Lab test environment, where each of the scenarios was repeated by varying the delay parameter. For each of these scenarios and network conditions the participants were asked for their subjective opinion (MOS). I obtained the statistics of the speaker alternation rate, mutual silence and double talk.

Finally, I analyzed the results so we can compare the delay impact as the one specified in the ITU recommendations.

1.2 Thesis Overview

Chapter 2 gives an overview of the relationship between QoS and QoE, Speech Quality Measurement, subjective and objective test methods and the ITU-T Recommendations related. Additionally the concept of conversational interactivity is described [5, Hammer 2006].

Chapter 3 describes the setup used for the tests, both hardware and software, test rooms and test persons.

Chapter 4 explains and documents the programs developed for the Voice Activity Detection and the calculation of relevant statistics about the interactivity that exists for each scenario and each network condition.

Chapter 5 describes the scenarios used and documents the correspondent results.

Chapter 6 gives a short summary of the thesis and some conclusions.

Chapter 2

Delay effects on speech quality

Voice over Internet Protocol (VoIP) can be viewed as a relatively new cost-saving technology for enterprises. For those commissioning and managing the data network transport of IP voice data over the local area network/wide area network (LAN/WAN), it may be seen like just another application to manage. The nature of the payload (voice, where there is no retransmission of time-sensitive packets) forces VoIP to mantain an entirely different effort in order to achieve a high quality of experience (QoE)

When a VoIP call is set up, speech is encapsulated in Real-Time Transport Protocol (RTP), which is encapsulated in User Datagram Protocol (UDP), both are transmitted in an IP frame. Each RTP packet contains a small portion of the voice conversation. The size of the voice sample is dependent of the codec used to compress the digital bit stream at any endpoint, such as an IP phone.

While a higher-bandwidth codec more accurately reproduces the analog input signal, it requires a higher bit rate, which generates more network traffic and reduces the network's overall call capacity. Using a lower bit rate codec sacrifices quality yet uses less bandwith.

Real Time Control Protocol (RTCP) allows the endpoints to communicate directly concerning the quality of the RTP packet stream. The control plane provides signaling protocols that perform such functions as register VoIP phones and connect phone calls.

Several common effects impair voice quality on a VoIP system. The test and measurement values important for managing QoE on a VoIP system are the following:

Delay: Because IP networks operate on statistical multiplexing technologies, latency in IP networks is usually higher than with analog transmission. Any delay in end-to-end transmission of voice from speaker to listener impedes voice quality. IP networks will have varying latency times over a single path depending upon the level of traffic on the network. In general, lower bit rate codec increases delay for VoIP calls.

Packet loss: This can occur in many ways. A router or switch queue may be full and cannot hold any more packets, causing arriving packets (segments of a voice transmission) to de discarded. Bit errors may exceed correctable levels, or a packet may be misrouted or exceed its time-to-live quota due to network topology changes or network congestion. In either case, packet loss harms VoIP quality.

Jitter: Packets that arrive at the destination at irregular intervals or out of sequence can make voice choppy and difficult to understand. Out-of-sequence packets often occur due to multiple routing paths to the same destination. If packets are out of sequence by only one or two sequence numbers, jitter buffers on receiving devices can place packets back into the order before voice playback. Packets unable to be placed into proper order are discarded by the receiving device, reducing voice quality.[10, Hestnes et al., 2003]

Applications such as email and file transfers tolerate packet delays and user retransmission of bad or missing packets to achieve error-free performance at the application level. Because VoIP cannot tolerate retransmission and demands priority routing of packets, it places more stringent requirements on IP data networks. Ultimately, as with video, VoIP service quality is determined subjectively by the end users.

For voice, unlike data, the key measures of quality are intelligibility and identification. Intelligibility is the ability to understand what is being said. Identification is the ability to recognize the voice of familiar callers.
Objective and subjective measurements exist to judge the performance and QoE

of VoIP services. Active tests, such as Perceptual Speech Quality Measurement (PSQM), Perceptual Evaluation of Speech Quality (PESQ) [11, ITU-T Rec. P.862], and Perceptual Analysis and Measurement System (PAMS) that use analog input signal, collect known voice samples across the network to a receiving endpoint, where a comparison analysis of the degraded sample with the original sequence is conducted. These are not tones, but rather actual prerecorded wav files available in different languages. This test requires two devices (one at each end) and is often used to evaluate the ability of the existing network to handle VoIP by generating and assessing calls.

2.1 Quality of Experience

A key challenge in the successful realization of network convergence is to ensure that all applications running over the network perform well, regardless of whether they are voice, video, data, real time, or non real time applications. The converged network must be able to efficiently carry all traffic types without degrading any of them, and it must meet the combined requirements for all services at a level equal to or even better than what the user has become accustomed to form a single service (voice only or data only) offering.

Addressing this challenge meet that at each level of network, performance parameters must meet the strictest requirements of all those defined for the individual services, while simultaneously delivering an overall acceptable quality of experience (QoE) that's include security, reliability and availability. QoE is the user's perception of how well a system, application or network interaction performs relative to their expectations, as well as how intuitively they can use an application or service to accomplish a task in a timely and efficient manner, without concern for the underlying network elements.

QoE and QoS are related, but they are not the same. For example, it is possible to have excellent QoS but poor QoE. Quality of Service (QoS) refers to a set of technologies that enable the network administrator to manage the effects of congestion.

Furthermore, while QoS is measured objectively, QoE is a subjective measurement that generally requires translation into quantitative data. QoE can be ob-

jectively quantified using standardized statistical procedures and various analysis methods.

Traditionally, providers have focused on QoS to ensure service performance. QoS involves such measurable parameters as service availability, delay, jitter, throughput, packet loss rate, bit error rate and signal to noise ratio. QoS helps operators to determine the levels of quality to use for different services, as well as to understand how to configure services, in order to differentiate them. At the same time, operators must balance this against the need to minimize cost and maximize link utilization.

With QoE, customer needs and expectations become central to product design and business processes. There is some engineering methodology to help drive QoE into the network planning and engineering processes. The objective of this methodology is to facilitate the selection of effective QoS mechanisms that satisfy the end user QoE of a given application , and is based on a top down approach, starting at the end-user level. A summary of the major steps of this methodology is the following:

Define service QoE performance metrics and targets.

Identify QoE contributing factors and dependencies:

- Impairments : delay, loss, jitter.
- Application decomposition.
- Client/server interaction
- Impairments : delay, loss, jitter.

Determine network architecture, QoS mechanisms and configuration:

- Define service levels guarantees.
- Network transformation phases, call scenarios.
- Echo canceller placement and transmission planning (voice)
- Nodal and end-to-end level: scheduling, policing, queue management.

Traffic engineering and resource allocation:

- Determine traffic demands, distribution and bottleneck links

- Budget allocation: delay, loss, jitter.
- Router resource and buffer dimensioning.
- Bandwith provisioning

Meet QoE targets:

- Yes: validated. Service QoE requirements are satisfied by the QoS enabled solution.
- No: back to determine network architecture, QoS mechanisms and configuration.

QoE is the user's perception of the performance of a device, a service or an application. User perception of quality is a fundamental determinant of acceptability and performance for any service platform. The design and engineering of telecommunications networks must take into account the perceptual , physical and cognitive abilities of the humans that will use them; otherwise, the performance of any service or application that runs on the network is likely to be unacceptable. Successful design requires a thorough understanding of the needs and constraints of the eventual users. QoE is measured and understood at the system level. For telecommunications networks , this means understanding the end-to-end performance.

2.2 Conversational Interactivity

To perform an analysis of conversational interactivity I employed the conversational model described in the next section. This model allowed me to obtain information on the status of the conversation and extract various parameters that were of great use in analyzing the impact of delay on speech quality.

2.2.1 The Conversation Model

The [12, ITU-T Rec. P.59] defines a two way conversation model including four different states, as shown in figure 2.1

States A and B represent the situation that either speaker A or speaker B is talking only. State M ("mutual silence") denotes the case that nobody talks at all, and state D ("double talk") reflects the situation that both speakers are talking simultaneously.

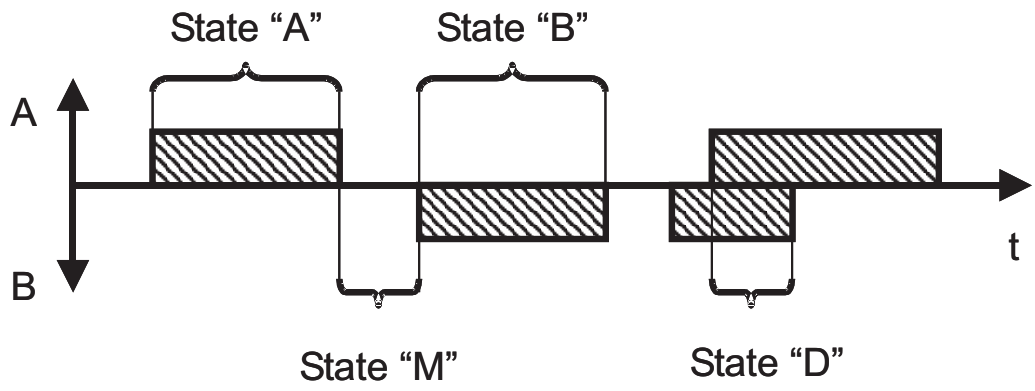


Figure 2.1: Conversational states

We can consider the following conversational parameters:

Speaker alternation rate. - Refers when the speaker changes. The correspondent transitions states could be A-M-B or B-M-A. The ”speaker alternation rate” (SAR) represents the number of speaker alternations per minute.

Pause. - It’s defined a phase of mutual silence. The transitions states could be A-M-A or B-M-B.

Interruption. - It’s defined as a phase where there is a double talk and after the speaker changes. The transitions states could be A-D-B or B-D-A.

Non interruption double talk. - It’s defined as the event of double talk occurring without ending up in an interruption (A-D-A or B-D-B).

These events are illustrated in figure 2.2

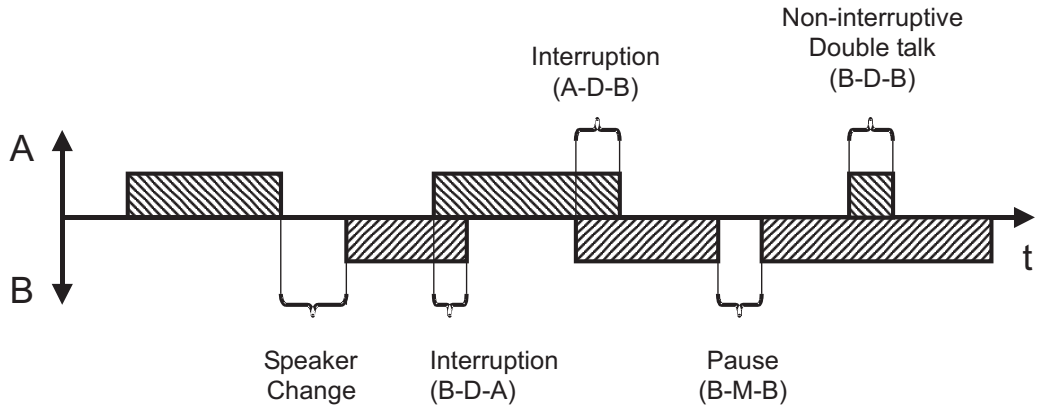


Figure 2.2: Conversational events

2.2.2 Models for conversational interactivity

[5, Hammer 2006] presents three models for conversational interactivity: the speaker alternation rate, a conversational temperature model and a model based on the entropy of speaker turns.

Speaker alternation rate. - Represents the number of speaker alternations per minute. A low SAR corresponds to low conversational interactivity and a high SAR corresponds to a highly interactive conversation. It can simply be calculated by counting the speaker alternations and dividing them by the duration of the call.

Conversational temperature. - The name of this concept is derived from the common speech, as people often denote their conversations in terms of "heat". A formula is developed to compute an estimated conversational temperature, which mainly depends on so called "sojourn times" within a particular state. Also a reference conversational temperature is defined which is chosen to be 21,5 C (room temperature).

Entropy. - It's based on a so called turn model and involves more than two interlocutors (multi-party conversations). Based upon's Shannon entropy formula, a formula for the entropy rate is derived that corresponds to the non predictability about which party is actually talking at which moment. The entropy rate depends on the mean overall turn duration and the turn probabilities for a certain number of speakers.

In this work, I used the Speaker Alternation Rate model in order to evaluate the delay impact on the speech quality.

2.3 Subjective Quality Test

The definition of the term "quality" according to [13, Jekosch, 2005] is "Quality is the result of the judgment of the perceived composition of an entity with respect to its desired composition". In this case, the entity to be judged is the speech transmission system. This system is used to establish and maintain an audio connection between two test persons. Each person has certain expectations of such a system, according to the experience. Thus, the term "quality" expresses the correlation between the user's expectations on one side and the actual state of the system as it is perceived on the other side.

Jekosch distinguishes quality elements and quality features. Quality elements are the characteristics of a system or service which are related to their design implementation or usage. Quality features represent the perceptual characteristics that contribute to the user's quality perception.

Perception based quality methods include all kinds of methods that are based on tests involving test persons either in listening only or conversational situations. Subjective tests are time consuming, expensive and require appropriate test facilities, even if they are performed on a relatively small scale. In order to reduce this effort and facilitate efficient and cost effective quality measurement, instrumental measurements methods have been developed. Further on this type of methods will be analyzed.

2.3.1 Speech quality measurement

Subjective quality measurement methods have been standardized in [1, ITU-T Rec. P.800.] This recommendation specifies three major methods:

1. *Conversation opinion tests.*
2. *Listening opinion tests*
3. *Interview and survey tests.*

The choice of an appropriate measurement method depends on the impairment to be tested. In listening only tests, degradations that directly impair the speech

signal can be measured. Conversational testing is used when we need to measure impairments that only occur in conversation situations, for example delay or echo. The results of subjective quality tests are typically presented as Mean Opinion Scores (MOS) which represent the mean ratings given by the tests subjects.

2.3.1.1 Conversation - opinion tests

In a conversation test, two test persons have a series of conversations over a real time telephone test system in a controlled laboratory environment. The subjects fulfill the tasks of a given conversation scenario. After each conversation, the subjects rate the quality of the connection they have been using on a five point scale from "Excellent" to "Bad" as can be seen in the next table.

Detailed description of the method, considerations and precautions are found in Annex A of the [1, ITU-T Rec. P.800.] This description includes:

> Section A.1. Describes the physical condition of the tests cabinets: dimensions, sound attenuation, etc. Details about the noise level in them are given, as well as recommendations on how to establish a connection and how monitoring is done.
>
> Section A.3. This section explains about the conversation task: "Every effort is to be made to ensure that conversations are purposeful, and the subjects have full opportunity to exploit the transmission capabilities of the test circuit". Also, it recommend that a test conversation should start and end in a natural way and it should not be interrupted during the task.
>
> Section A.4. Refers to the test procedure . It include the following items : a) Eligibility of subjects. Subjects taking part in the conversation tests are chosen at random from the normal telephone using population , with the following criteria :
>
> – a) Eligibility of subjects. Subjects taking part in the conversation tests are chosen at random from the normal telephone using population , with the following criteria :
>
> > Subjects must not be experts in quality assessment of telephone circuits or corresponding fields.
> >
> > Subjects must not have attended any subjective test in the previous half year, as well as no conversation test for at least one year.

– b) Opinion scale:

The experimenter allocates the following values to the scores :

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

Table 2.1: MOS table

and all further statistical processing is performed in terms of these numbers. Additional details about the instructions to subjects and how to manage the data collection and the treatment of results are included in this section.

2.3.1.2 Listening opinion tests

In this kind of tests, the subject listen to a series of speech samples and rate the quality based on an appropriate rating scale. This test method is highly suitable for investigating the effects of different speech coding algorithms, but it isn't expected to reach the same standard of realism as conversation tests, and the restrictions are therefore less severe in some respects.

2.3.1.3 Interview and survey tests

When the importance of study warrants it, transmission quality can be determined by "service observations". Recommended ways of performing these, including the questions to be asked when interviewing customers are given in [14, ITU-T Rec. P.82]. To maintain a high degree of precision a total of al least 100 interviews per condition is required. A disadvantage of the service observation method for many purposes is that little control is possible over the detailed characteristics of the telephone connections being tested.

In this work, I used conversation testing method because is more appropiated to measure the delay impairment.

2.3.2 Scenarios

In the following sections I describe the test scenarios more common associated to the speech quality assessment:

a) Short Conversation Test (SCT). This scenario proposes to establish a natural and balanced conversation of about three minutes. Natural means that test persons should act like they would in an everyday situation. Conversational Parameters should be balanced for example by means of equal talking/listening times. The SCT represents today's standard scenarios in conversational speech quality assessment and is based on tasks like ordering pizza or booking an hotel room.

b) Interactive Short Conversation Test (iSCT). This scenario is based upon the "rapid exchange of numerical and lexical data". One test person has to obtain missing data items from its partner that has the accordant information at hand, and vice versa. Another example of this type of test is the Random Number Verification (RNV). This scenario requires the rapid verification of a given set of random numbers. The test persons are asked to alternately verify the numbers either in rows or in columns. It is expected to be highly interactive and to yield high impact of transmission delay on perceived quality.

c) Asymmetric Short Conversation (aSCT). It is similar to iSCT, but in the aSCT tasks the called person is given all of the information, while the calling person needs to request it. Thus, the structure of the resulting conversations is expected to be asymmetric by means of the speech activity of the two participants.

d) Free Conversation. The free conversation scenarios results in "everyday" conversation of about seven minutes based on given topics. The structure of the conversation is not strictly predetermined by a given task, but rather driven by the conversation behavior of the test subjects. The scenario may depend heavily on the test persons themselves (differences in age, personality) and how interested they are in the proposed topic.

2.3.3 Rating Scales

Different rating scales methods has been standardized in [1, ITU-T Rec. P.800.], Annex B,C,D,E. These rating scales are commonly used in the field of subjective quality test. The following sections summarize an overview of different rating scales.

1. - Absolute Category Rating (ACR). [1, Annex B ITU-T Rec. P.800.] describe the following items:

- B.1. Is about the procedure of source recordings. In several sections it develop different details about the recording environment , sending system alternatives of recording system, characteristics of speech material, recording procedure and behavior of the talkers.
- B.2. This section details the selection of circuit conditions, listening levels and reference conditions.
- B.4. It describes the listening test procedure, characteristics of listening environment and listening system. The recommendations about the listeners are the following:

  They are chosen at random from the normal telephone using population.

  They have not been directly in work connected with assessment of the performance of telephone circuits, or related work such as speech coding.

  They have not participated in any subjective test whatever for at least the previous six months, and not in any listening opinion test for at least one year.

  They have never heard the same sentences lists before.

  The opinion scales recommended maps the subjective opinion to a numerical value and it's included the following:

  A.- Listening quality scale:

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

Table 2.2: Qualities of the speech

The quantity evaluated from the scores (mean opinion score) is represented by the symbol MOS_{LEC}

B.- Listening effort scale:

Effort required to understand the meaning of sentences	Score
No effort required	5
No appreciable effort required	4
Moderate effort required	3
Considerable effort required	2
No meaning understood with any flexible effort	1

Table 2.3: Efforts table

The quantity evaluated from the score is represented by the symbol MOS_{LP}

C.- Loudness preference scale:

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

Table 2.4: Loudness table

The quantity evaluated from the score is represented by the symbol

- *2. - Mean Opinion Score (MOS).* This can be computed from such a scale as the mean value of the subjective opinion. The following identifiers are recommended to be used together with the abbreviation MOS in order to distinguish the area of application, where LQ refers to Listening Quality, CQ refers to Conversational Quality, S refers to Subjective, O refers to Objective and E refers to Estimated.

	Listening only	Conversational
Subjective	MOS-LQS	MOS-CQS
Objective	MOS-LQO	MOS-CQO
Estimated	MOS-LQE	MOS-CQE

Table 2.5: MOS table

Commonly, the MOS is represented as a function of impairment. This representation allows quick determination of the average subjective opinion for a given impairment factor.

- *3. - Degradation Category Rating (DCR).* [1, Annex D, ITU-T Rec. P.800] describes this rating scale. This scale consists of five grades that details

the level of impairment, form Imperceptible (5) to Very Annoying (1). The DCR method is used to distinguish among good quality transmission system for which the ACR method lacks sensitivity. In DCR test, the degradation of samples that passed through the system under test is rated against a high quality reference.

Impairment	Score
Imperceptible	5
Perceptible, but not annoying	4
Slightly annoying	3
Annoying	2
Very annoying	1

Table 2.6: DCR table

– *4. - Comparison Category Rating (CCR).* [1, Annex E, ITU-T Rec. P.800] details this kind of rating. While in the DCR method the reference sample is always to be presented first, in the CCR method the order of the processed and reference sample is randomly chosen for each trial. In the half of the trials, the reference sample is presented first, and in the rest of the trials, the processed signal is presented first. After each trial, the test persons are required to rate the quality of the second sample in comparison with the quality of the first using the following scale:

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

Table 2.7: Loudness table

The advantage of the CCR method over the DCR method is the possibility to measure the impact of speech processing that either impairs or improves the quality.

From the point of view of the subjective quality test, in this work I used the MOS rating scale.

2.3.4 Interactive audiovisual quality

As an extension of methods defined in [1, 2, ITU-T Rec. P800 Annex A and ITU-T Rec. P.920] it has been defined interactive evaluation methods for quantifying the impact of transmission delay and transmission impairments on point-to-point or multipoint audiovisual communications. This methodology is based upon conversation opinion tests.

Experimental design: The ITU-T Rec. P.920 [2] is based on an active talker conversation assessment, in order to quantify the impact of factors such as transmission delay, that may affect the ability to conduct an interactive communication. In conversational opinion tests, it is desired to minimize the artificiality of the environment. But, at the same time, it is necessary to invoke some methods to stimulate interactive communication utilizing the conditions which are being evaluated. For general applications, the following guidelines are provided for design task-based tests :

- The task should be designed such that, during their conversation, the subjects primarily maintain their attention on the audiovisual terminal.
- The task must have sufficient face value, that is, it must resemble real life audiovisual communication to a sufficient degree.
- The task must yield reproducible quantitative results that represent adequate measures of communication efficiency.

It is recommended that the task is sufficiently rewarding for the subjects. In this case, the subjects learn the task faster and they are less susceptible to fatigue and loss of motivation.
About the test conditions, it's recommended that at least one transmission impairment factor is likely to be evaluated in a test, in addition to a reference condition where the impact of such factor is minimum. It's necessary to consider that conversational tests are time consuming, hence the total number of conditions has to be reasonably constrained in order to minimize participant fatigue and maximize experimental accuracy. This requirement should be balanced against the need to ensure that the duration of each conversation/-condition is at least five minutes long.
After the completion of each condition, each participant should be separately solicited for his opinion. Different scales can be used to evaluate the audiovisual terminal performance. Examples of scales that may be used are the following:

- Overall Audiovisual Quality, Video Quality and Audio Quality are generally assessed using the categories: Excellent, Good, Fair, Poor and Bad.
- Effort needed to interrupt is generally assessed using the categories: No Effort, Minor Effort, Moderate Effort, Considerable Effort and Extreme Effort.
- Communications difficulty and acceptability of communications are generally assessed using a binary choice: Yes or No.

In [2, Appendix I ITU-T Rec. P.920] describe examples tasks for stimulating a conversation. These are the following:

- *Name-Guessing:* a question-answer game which follows precisely a defined protocol. One subject has information about either a well known person or a certain brand. The other subject has to guess the name.
- *Story-Comparison:* before the tests, the subjects have to read and memorize a short story- There is a certain amount of differences between their stories which have to be detected as quickly as possible.
- *Picture-Comparison:* subjects are given pictures which they have to memorize. The goal of this task is to jointly make a decision whether they were given identical pictures or not.

Appendix I, section 3, proposes several scenarios to force test persons to use the video channel: building blocks and object description.

Chapter 3

Experimental Environment

3.1 Setup Description

3.1.1 Hardware

I used two laptops with headsets for the test users each one in a different room. Another computer was used to manage the system from the Control room. Also it was used a network emulator to select the net conditions for each test in the VoIP network.

Laptop	HP 8530p model number: FU457EA
Screen Size	15.4 inch
Operating System	Windows XP Professional
VoIP Clients	Mizu Softphone
Headset	Sennheiser PC 21 (monaural)
Sound card	Phase 26 Terratec Producer
Mixer	Xenyx 802
Microphones	Bahringer Ecnm8000

Table 3.1: Characterization of the user terminal instrumentation

3.1.2 Software

I used a Mizu Softphone client for stablishing the commmunication between the participants. Mizu Softphone is a professional VoIP softphone based on

the open standard SIP protocol with an easy to use interface. Also a Network Emulator provided by FTW to set the delay conditions for each conversation was used. This Emulator used Ubuntu 8.10 Server version. As communication server I used Asterisk with the Ubuntu 8.10 Server version.

3.1.2.1 End-to-End Delay

The delay settings were controlled by the control PC which used Telecommander to manage it. Telecommander is a proprietary solution of FTW to handle various network conditions. It allows to select the delay for each condition and saves automatically the Mean Opinion Score of the participants after sending a prompter with the questions to participants laptop after each test. Telecommander uses Python 2.6 as software.

3.2 Test Rooms

The tests were conducted in the i:lab and the library of FTW facilities. Each participant had its own laptop with headset in the i:lab (participants 'A') and the library (participants 'B'). Another room (control room) was used to control the tests. Participants were given instructions from a microphone in the control room by a speaker located in each room. Library and i:lab room were completely isolated between themselves and also from outside noise. Both rooms were prepared to perform real conversation in a quiet environment.

3.3 Test persons

Test were conducted over three sessions in March 2010, days Friday 19, Thursday 25 and Friday 26. 34 people participated in them. Almost all of them have never participated in a test. Only German speaker participants were familiarized with tests. Tests 1 and 7 were cancelled due to technical problems with microphones and emulator. Tests were conducted in three languages: Spanish (12), German (4) and Hungarian (1). About a 44 per cent of participants were engineers and 32 per cent of them were female. The age ranged from 18 to 28 years with an average age of 23.79 years.

Participants were rewarded for their participation with a EUR 15,- voucher in Saturn shops.

No. of pair	Age	Field of study	Sex	Language
2-A	23	Industrial Eng.	M	Spanish
2-B	22	Medicine	M	Spanish
3-A	23	Industrial Eng.	F	Spanish
3-B	26	Industrial Eng.	M	Spanish
4-A	21	Biotechnology	M	Spanish
4-B	24	Informatics Eng.	M	Spanish
5-A	24	Informatics Eng.	M	Spanish
5-B	21	Psychology	M	Spanish
6-A	22	Telecomm. Eng	M	Spanish
6-B	22	Telecomm. Eng	M	Spanish
8-A	25	Telecomm. Eng	M	Spanish
8-B	24	Law	M	Spanish
9-A	22	Chemical Eng.	F	Spanish
9-B	26	Industrial Eng.	F	Spanish
10-A	21	Business	M	Spanish
10-B	21	Architecture	M	Spanish
11-A	22	Business	M	Spanish
11-B	28	Informatics Eng.	F	Spanish
12-A	22	Medicine	M	Spanish
12-B	24	Informatics Eng.	F	Spanish
13-A	26	Telecomm. Eng	M	Spanish
13-B	27	Telecomm. Eng	M	Spanish
14-A	19	German	F	Hungarian
14-B	18	Communication	F	Hungarian
15-A	23	Business economics	M	German
15-B	36	Editorial journalist	M	German
16-A	23	Telecomm. Eng	M	Spanish
16-B	25	Translation	F	Spanish
17-A	31	Ecology	M	German
17-B	24	Human Ethology	F	German
18-A	21	Zoology	F	German
18-B	22	Moleculare Biology	F	German
19-A	26	Genetics	M	German
19-B	25	Teacher-training Course	M	German

Table 3.2: Test participants

Chapter 4

Operating procedures and algorithms

Having described the test environment in the previous chapter, I explain in this chapter the fundamentals, procedures and programs developed that will be used.

The procedures and programs described in this chapter permit to get different conversational parameters. The analysis of these metrics enables us to obtain a series of conclusions about the delay impact on the speech quality for different types of the scenarios and conditions tested.

4.1 General procedure

Using the test environment described in the preceding chapter, each pair of participants in the evaluation of the scenarios will use the script for the different types of test scenarios defined:

- 1.-Short Conversation Task (SCT)
- 2.- Interactive Short Conversation Task (ISCT)
- 3.- RNV

Each of the above scenarios will be used with five different delay conditions that will be programmed in the corresponding network emulator. In this way we obtain 15 different situations for each participant pair. The block diagram of the general procedure used is described in Figure 4.1. The procedure that supports the automation of the process is included in Appendix A.

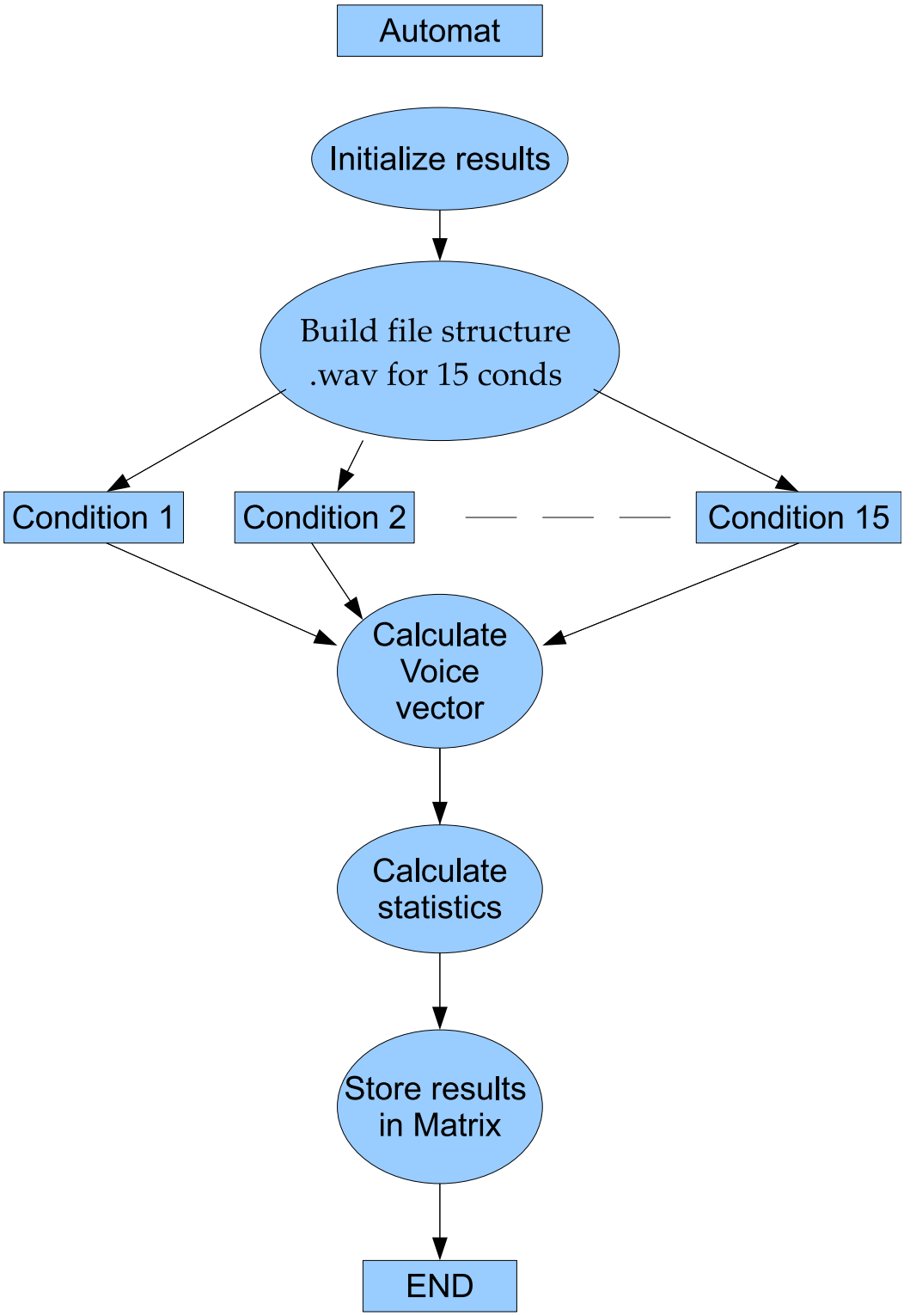


Figure 4.1: Automat block diagram

The steps included in this procedure are the following :

- 1 .- Construction of the file structure .wav corresponding to the recordings made for each of the members of the participating pairs.
  After stablishing the communication between each pair of participants, there will be a synchronization system that allows, almost simultaneously, to start the recording of the dialogue in the .wav file specified in the corresponding stage on the hard drive of each of the PCs used by members of the pair. The synchronization procedure of the recordings of each participating partner is programmed in Tool Command Language and is included in the Appendix C. The main scheme of the synchronizer is that the control computer sends a socket to both participant computers with the command of starting to record at the same time from the microphones in .wav files in each computer. Therefore there are 3 programmes, one in each computer of the system.
  Once completed the cycle of test scenarios for each participant pair, we will have 30 files .wav, 15 for the 15 different conditions in which it has participated each partner.
- 2 .- Voice Activity Detection.
  After constructing the structure of the .wav files in a subdirectory for each one of the participating pairs, we use a specific algorithm developed to detect the existence of voice or silence. This program builds a energy vector where each component of this correspond to the energy levels of the signal stored in the .wav file. Subsequently, it calculates a certain energy threshold . The calculation of this threshold is based on an average calculated over a period of initial silence required and maximum energy level. The calculated energy thresholds are used in an algorithm to be described later and allows us to build for each wav file, a new vector where each segment of 5 ms of signal is indicated with "1" if voice is detected and with "0" if it corresponds to silence.

- 3 .- Calculation of statistical parameters and storing the result in a matrix. This program uses as input the voice detection vectors obtained in the previous stage, for each pair of participants and for each test condition. The objective of this program is to obtain the following statistics:

  A) Time of mutual silence (NMS): is the sum of the number of existing segments in the corresponding interval in which the two speakers have been silent.

  B) Time of Double Talk (NDT): is the sum of the number of existing segments in the corresponding interval in which the two speakers have been talking simultaneously.

  C) Number of Speaker Alternation Rate (NAT): is the sum of the number of existing segments in the corresponding interval in which the two speakers have alternated in the use of the word.
  Finally we obtain the cumulative total for the vector and corresponding recording for the above statistics:

  Total number of mutual silence

  Total number of double talk

  Total Speaker Alternation Rate

As example, here's the 9th pair and the 14th condition, a graphical representation of the files used in the general procedure described in this paragraph.

1 .- .wav files for A and B members of the pair.

2 .- Energy Vectors obtained for each member and their respective thresholds of energy used (ITL, ITU) (blue) and Voice Activity Detection vectors for members A and B (red).

3 .- Statistical parameters.

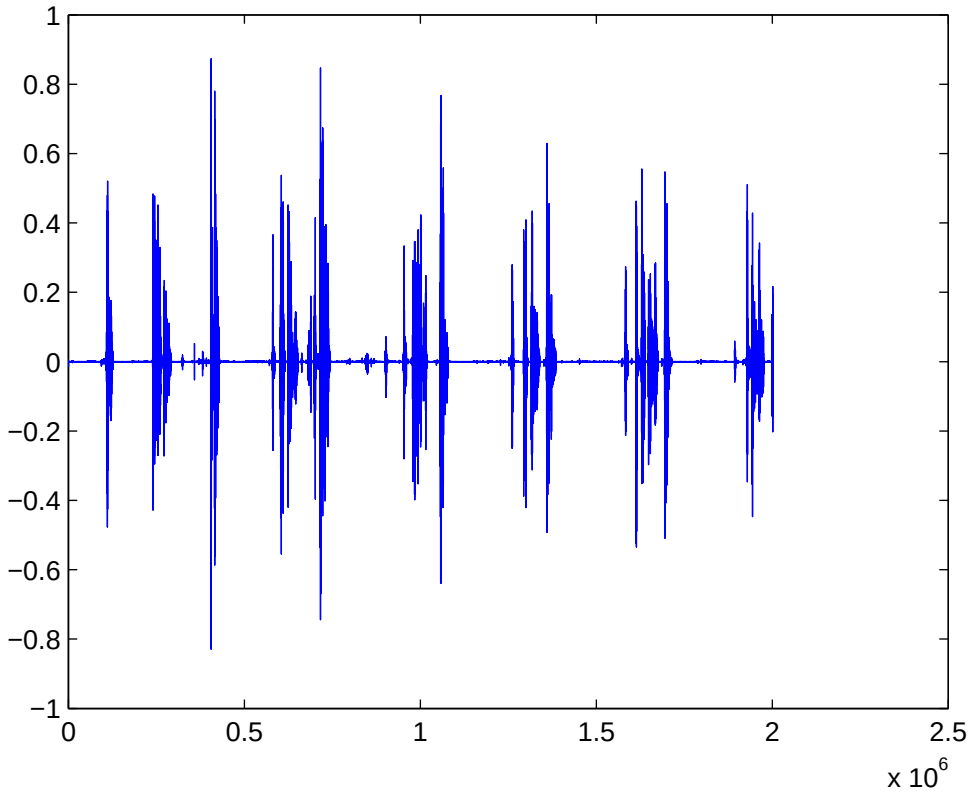


Figure 4.2: .wav file A

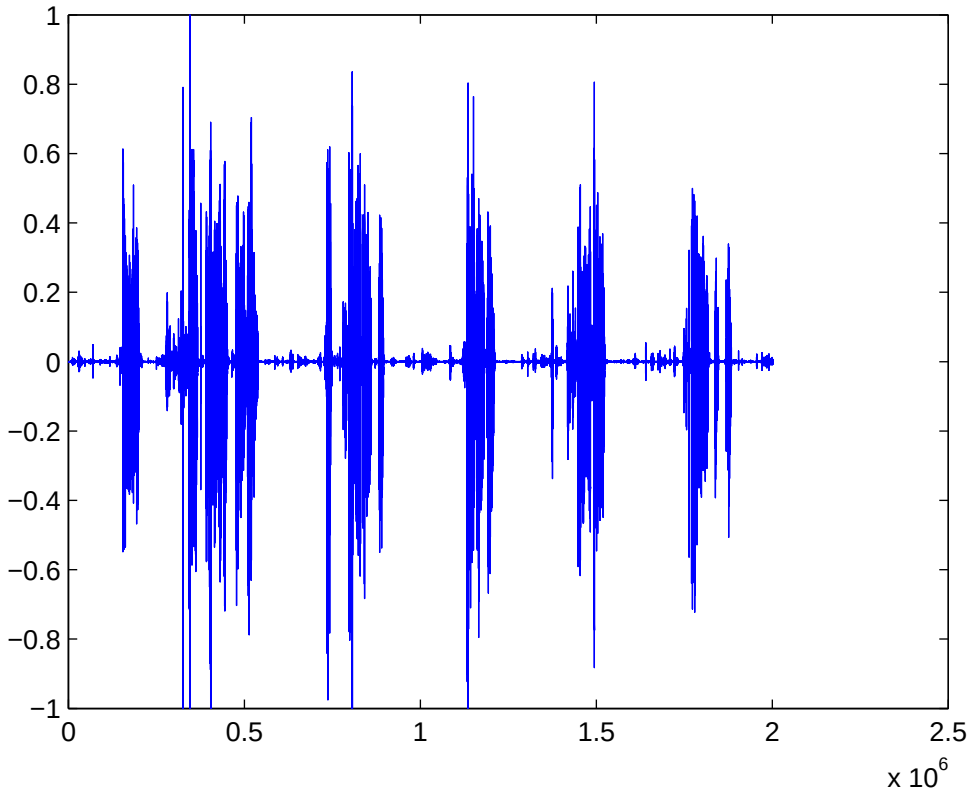


Figure 4.3: .wav file B

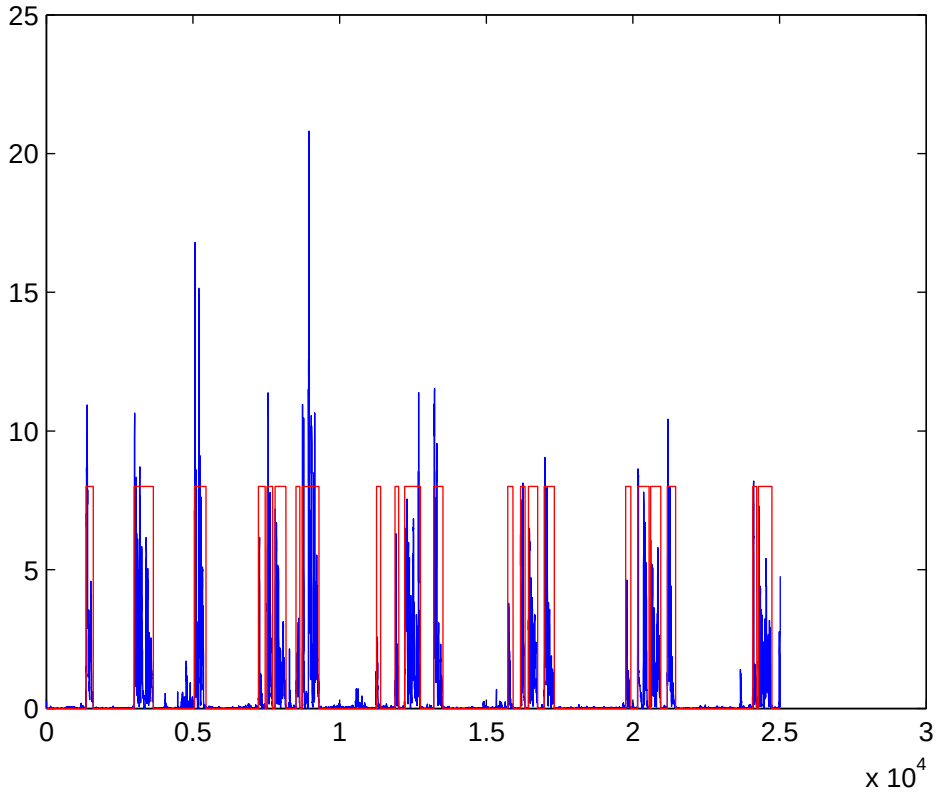


Figure 4.4: Energy Vector and VAD. Speaker A

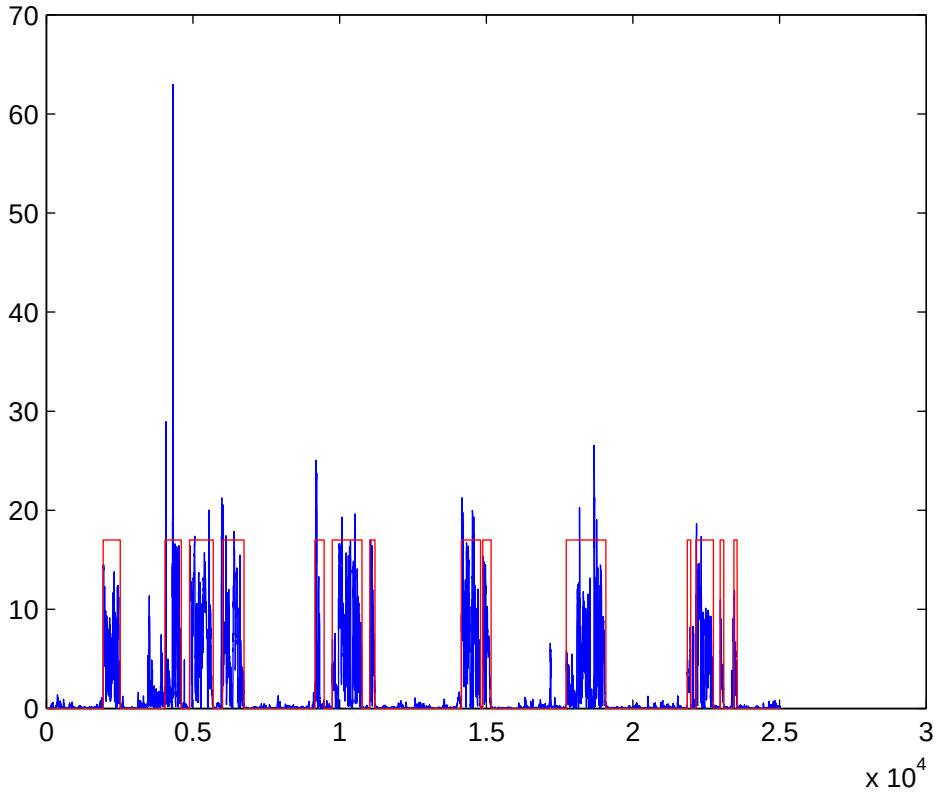


Figure 4.5: Energy Vector and VAD. Speaker B

In this case the metrics obtained are the following:

- Total number of mutual silence: 19
- Total number of double talk : 17
- Total Speaker Alternation Rate: 26

4.2 Voice activity detection algorithm

According to previous section, this program receives as input a .wav file and returns a vector which indicates for each segment that divides the file, '1 'if there is voice and '0' if there is silence.
The block diagram of the program developed is described in Figure 4.6 and the full program listing is included in Appendix C . The initial part of the program includes the calculation of the energy vector and the energy thresholds. After, this parameters are used to identify the segments corresponding to speech or silence. In this step, I use an algorithm proposed by Rabiner and Sambur.

This program requires in the recorded .wav file an initial period of silence of at least 100 ms in order to obtain several parameters related with this condition. This program includes a series of calculations and algorithms whose sequence is explained below:

- 1 .- It uses a sampling frequency of 16 kHz and calculates the number of segments of 5 ms that contains the complete file, and the number of segments for the initial silent period (100 ms). Initializes the Voice Activity Detection vector (vectorVAD) so that initially considers all segments correspond to silence ('0 ').
- 2 .- Calculates the Energy Vector for every one segment that has divided the .wav file. The value of each component of the Energy Vector is defined as the sum of the absolute values of the samples taken from the corresponding waveform for 10 ms.
- 3.- Calculate the following parameters:

  IMN: mean silence energy (noise energy)

  IMX: maximum energy

  ITL: Lower energy threshold

  ITU: Upper energy threshold

  The calculation is as follows:

1. I1= 0.03*(IMX-IMN)+IMN
2. I2= 4*IMIN
3. ITL= min (I1,I2)
4. ITU= 5*ITL

– 4.- Defines the following parameters:

  NsegsITLVoice=90. Indicates that the necessary (but not sufficient) condition to identify the beginning of voice is to find 90 consecutive segments (0,45 seconds) whose energy values exceed the ITL threshold.

  NsegsITUVoice=20. Indicates that the necessary (but not sufficient) condition to identify the beginning of voice is to find 20 consecutive segments (0,1 seconds) whose energy values exceed the ITU threshold.

  NsegsITLSilence= 70. Indicates that the necessary condition to identify the beginning of silence is to find 70 consecutive segments (0,35 seconds) whose energy values are lower than the ITL threshold.

– 5.- Makes a round through every element of the Energy Vector previously calculated to identify segments of voice or silence respectively. The algorithm developed include the following steps :

  A) Checks for at least 90 consecutive segments (0.45 secs) whose values exceed the threshold energy ITL. If the condition is right, it stores in the ITLs variable the index number corresponding to the first segment where ITL threshold is exceeded.

  B) Checks if, simultaneously the fulfillment of the above condition, there are at least twenty consecutive segments (0.1 secs) whose values exceed the ITU threshold energy. If both conditions are met, it will be considered that the variable ITLs stores definitely the index for the first segment of a voice interval.

  C) From the above condition, it must be identified where to end the voice interval whose beginning it was identified before. For this I will check if there are at least 70 consecutive segments (0.35 secs) whose energy values are below the threshold ITL. If the condition is satisfied, it stores in ITLf variable the index number corresponding to the end of the interval of voice.

  D) It is registered in vectorVAD the voice interval, writing "1" on all elements of it, included among the index identified in the variables ITLs and ITLf.

E) Once that is complete the above analysis for all elements of the vector energy are stored as a result the following variables:

vectorVAD that records the voice detection

Energy vector

ITL and ITU energy thresholds

4.3 Calculation of statistical parameters and storing results

The objective of this program is to calculate the statistical parameters that I use as the basis for the corresponding analysis.

This program uses as input the pair of VAD vectors in which it has been recorded the presence of voice or silence for each one of the pairs and test conditions.

The block diagram of the program developed is described in Figure 4.9 and the full program listing is included in Appendix A.

The sequence of algorithms and calculations of this program is as follows:

- 1.- It Identifies the shortest vector length and uses it as a basis for further calculations.
- 2.- It detects which of the two participants of the pair is the first speaker to use as reference for the subsequent calculation of the alternance.
- 3.- It divides the length of the vector calculated at intervals of 10 seconds in order to calculate the following statistical parameters.

  A) Time of mutual silence: is calculated as the sum of the number of existing segments in the corresponding interval of 10 seconds in which the two speakers have been silent.

  B) Time of double talk: is calculated as the addition of the number of existing segments in the corresponding interval of 10 seconds in which two speakers have been talking simultaneously.

  C) Number of speaker alternation rate: is calculated as the sum of the number of existing segments in the range of 10 seconds for the two speakers who have alternated in the use of the word. Described calculated values for each interval of 10 seconds are stored respectively in the following vectors:

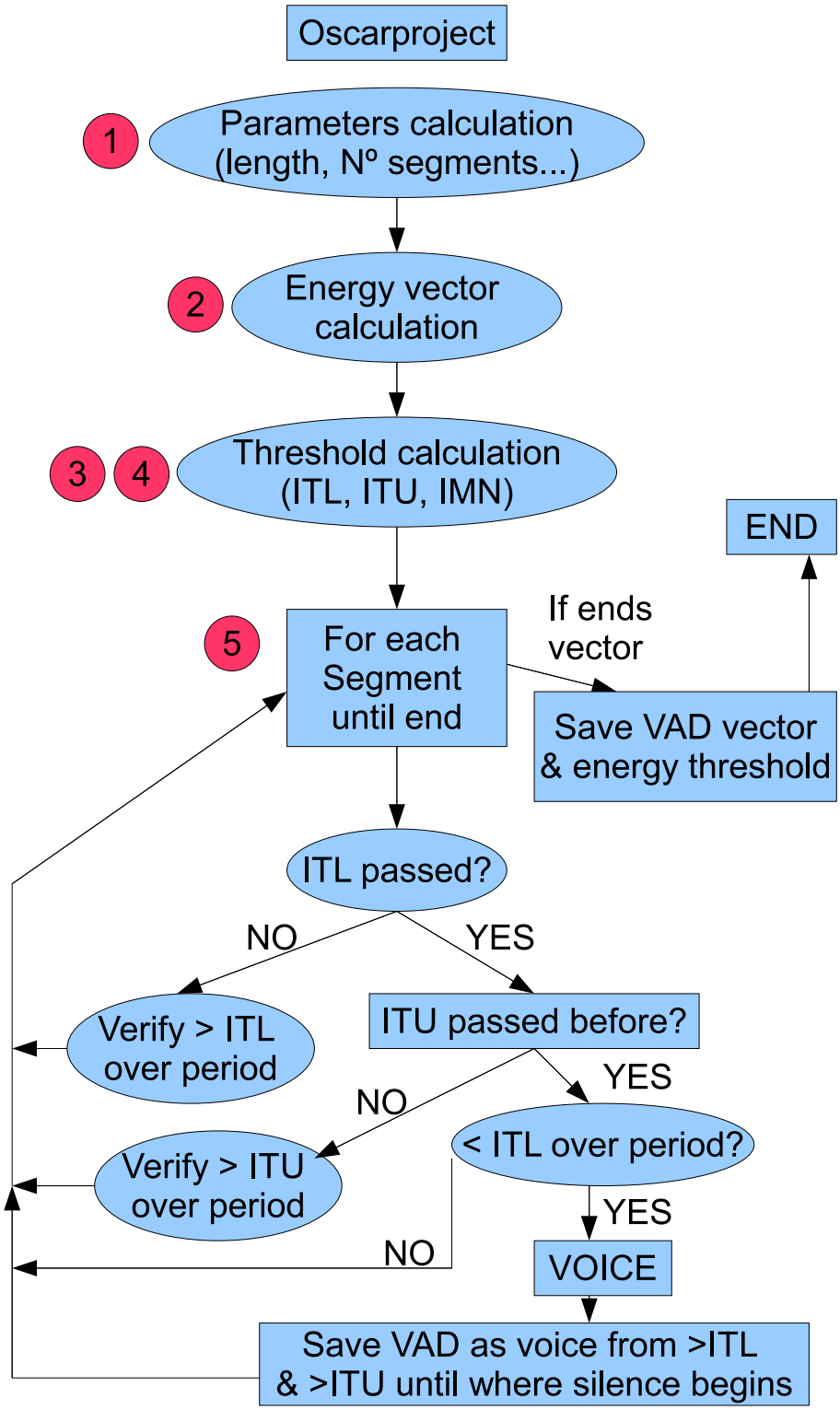


Figure 4.6: Main project program

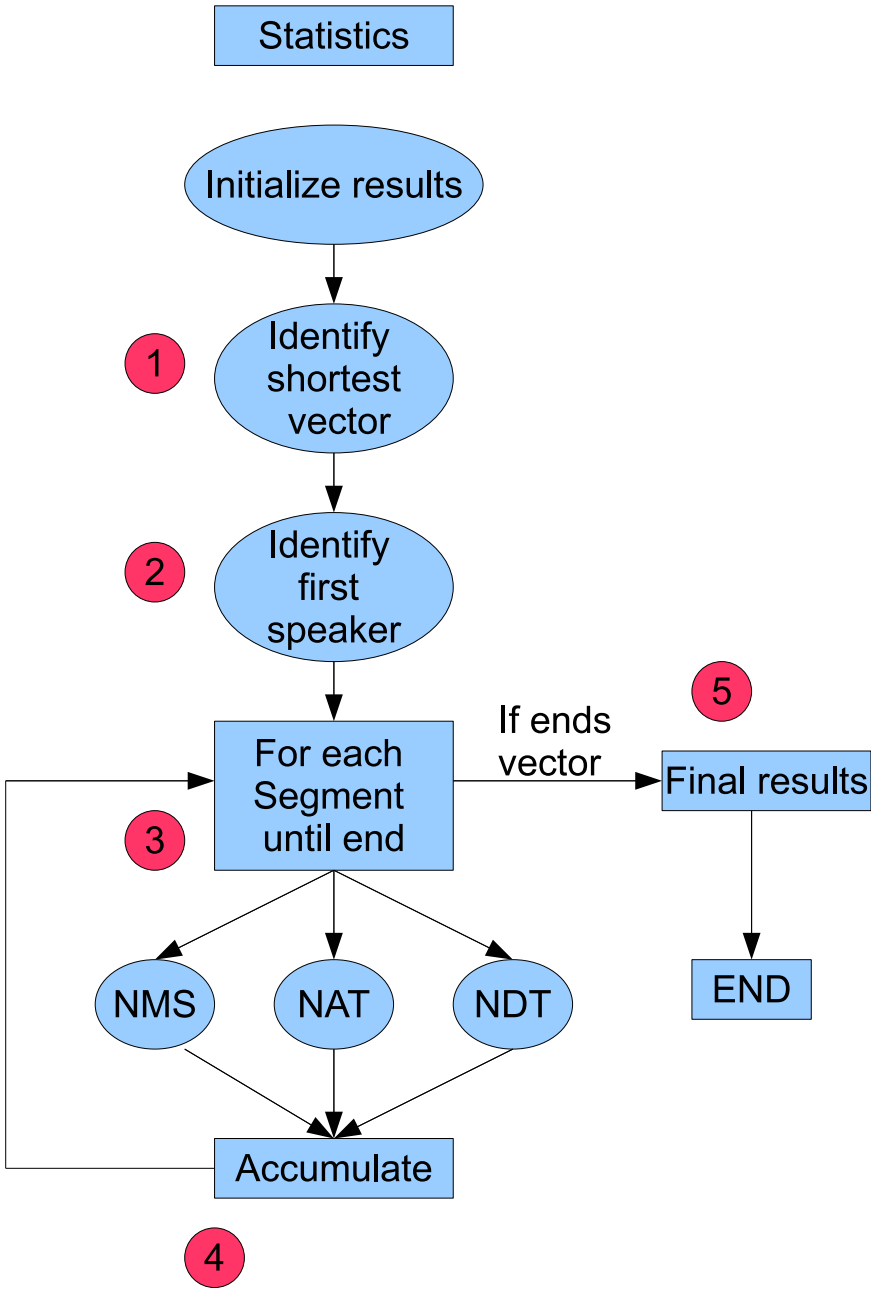


Figure 4.7: Statistics program

vectorMS: mutual silence

vectorDT: double talk

vectorAT: speaker alternation rate

- 4 .- After the calculation explained above for all intervals of 10 seconds, calculate the following parameters:

  NMSTOTAL: cumulative total number of mutual silence.

  NDTTOTAL: cumulative total number of double talk.

  NATTOTAL: cumulative total number of speaker alternation rate.

- 5 .- It is stored as a result the following variables: vectorMS, vectorDT, vectorAT, NMSTOTAL, NDTTOTAL, NATTOTAL.

Chapter 5

Evaluation of test scenarios

In this work, one of the major objectives is to build an experimental environment that permits to measure the delay impact on human to human mediated interaction. These measurements will be analyzed and this analysis will permit to establish the adequate thresholds. Beside that, I will compare the results obtained with the indications made by the ITU-T in several recommendations, especially in the [7, ITU-T Rec. G.114].

The perception of delay strongly depends on the conversational situation. If the purpose of a conversational test is to assess the network quality as it is perceived by an end user in a real telephony situation, the test scenarios should be close to that situation. This is because conversational situations are strongly influenced by the degree of interaction between the participants. For example, in stressful situations where lots of information must be transmitted in very short time, the subjects become more sensitive to delay effects as they more often may interrupt each other unwillingly (double talk).

Different conversational tests evoke different conversational interactivity. It's possible to measure this parameter evaluating the number of role (listener-talker) swaps per minute between the conversation participants. This parameter is denoted as Speaker Alternation Rate (SAR).

Conversation tests are a much more appropriate way of assessing communication efficiency, as the use of the conversational scenarios is more natural compared to the listening only tests. It's necessary to design conversation tests in such a way that they put the judging test subject into a highly natural situation. The key to achieve this is the design of an appropriate test scenario. The development of new conversation test scenarios have to achieve the reduction

of the conversation time, and consistently the associated cost, maintaining a natural and balanced dialog.

5.1 Requirements for conversation test scenarios

The requirements for conversation test scenarios are multiple. We can divide them in the following categories:

- **Naturalness.** Refers to the aspects of dependence subject :

  Natural communication task for the conversation (everyday situation, meaningful use of the telephone, no role play).

  Natural beginning and ending of each conversation

  Scenario should not distract the subject's attention too strongly.
- **Balance.** Refers aspect regarding the course of the dialog :

  No stable roles sender/receiver.

  Short monologues on both sides.

  Parts with various turn-takings for both sides.

  Provokes double-talk

  Balanced duration of each side.

  Conversation as short as possible.
- **Comparability between scenarios for different conversations :**

  Comparable instructions for all conversations.

  Standardized dialog structure.

  Constant dialog duration.
- **Clarity regarding the test instruction:** easy to handle instructions.

Obviously, several of these requirements exclude each other, so only a compromise can be reached with real scenarios.

Based on the requirements given above, some authors proposed the dialog structure that is illustrated in figure 5.1. This proposal includes parts with monologues, other where speakers speak in turns and some parts which should evoke double talk.

In the following sections I will describe the characteristics of different groups of test scenarios that I have been developed for this work. The general objective

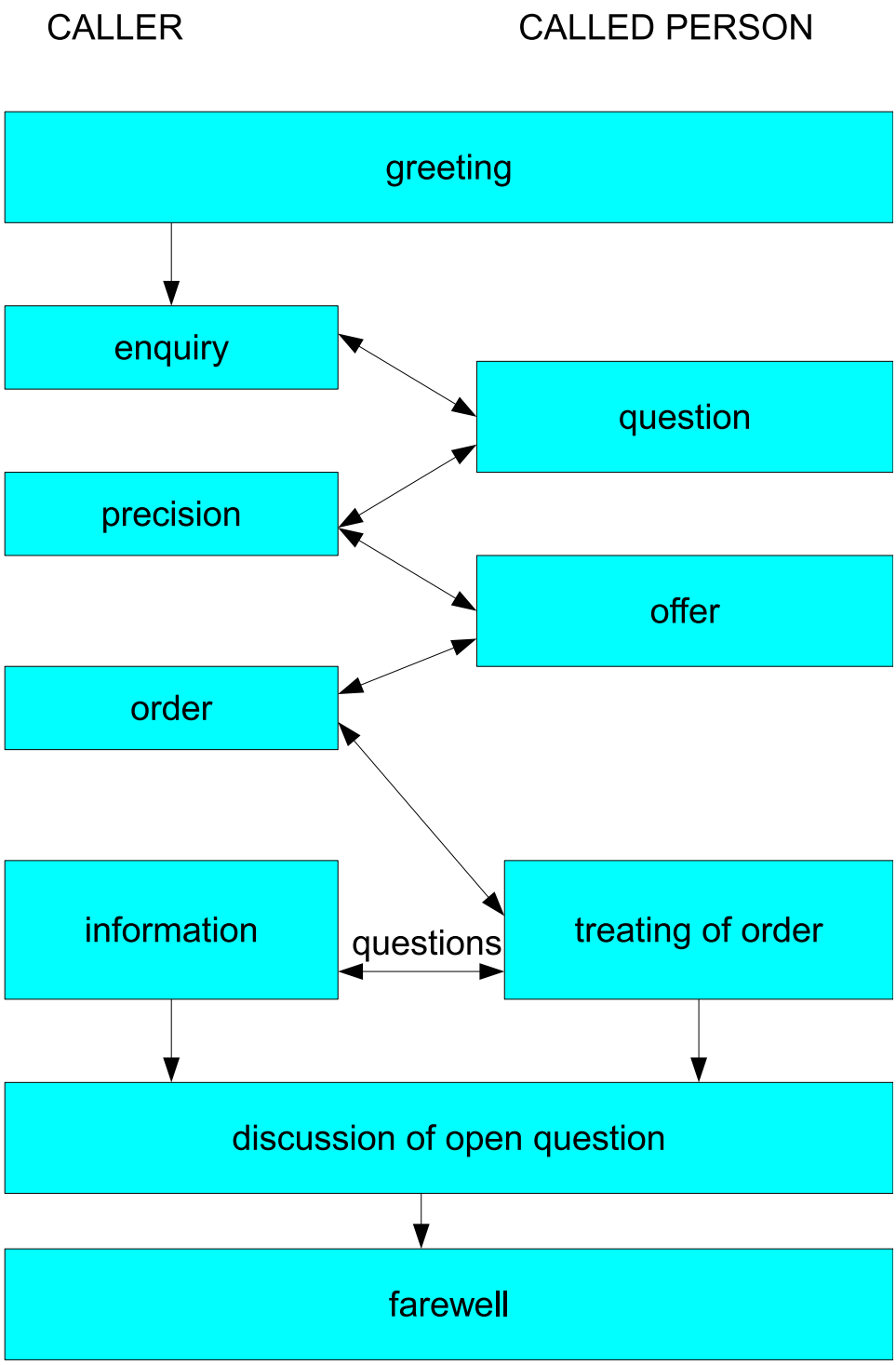


Figure 5.1: Course of conversation

of these test scenarios is to measure the delay impact on speech quality.
I have used several types of delay, each one with a determinate test scenario of each group.
At each participant's side, the conversation has been recorded on a PC. The result obtained for each delay and test scenario are the following:

- Speaker Alternation Rate.
- Mutual Silence
- Double Talk
- MOS_{CQS}. This opinion rating will be obtained interviewing the participants immediately after each test condition. In the next sections, I will explain for each scenario its contents and the test procedure applied. At the end of this chapter, I will describe the results of the different tests.

5.2 Short Conversation Test (SCT)

The SCT represents today's standard scenario in conversational speech quality assessment and is based on tasks like ordering a pizza or booking a hotel room. The SCTs result in natural, balanced conversations of about 2 minutes. Previous tests suggest that SCTs do not lead to sufficient conversational interactivity to generate significant impact of delay on perceptual quality. Specifically in this tests the used models were:

1. Pizza order
2. Flight booking
3. Holiday booking
4. Rent a car
5. Vegetable box order

5.2.1 Test procedure

One of the participants had the role of the caller and the other was the called. In order to have a more balanced results the role was always changed after each SCT test. So, in test SCT_1 participant 'A' was the caller and 'B' was the called, and in test SCT_2 participant 'B' was the caller and 'A' the called, for test SCT_3 roles were changed again and so.

5.3 Interactive Short Conversation Test (iSCT)

The iSCT scenarios should lead to comparable and balanced conversations of higher interactivity compared to the standard Short Conversation Test (SCT) scenarios. iSCT Scenarios were about date exchange between the participants. These tests focused on:

1. Temperatures and humidity data exchange in some austrian cities
2. Temperatures and snow depth in some austrian ski resorts
3. Shares variation data exchange between two stock exchanges
4. Music charts data exchange between two radio stations
5. Temperature of land and sea in this and last months in some resorts data exchange between two travel agencies

5.3.1 Test procedure

The plan of role changing between participants was the same that in SCT.

5.4 Random Number Verification (RNV)

This scenario requires the rapid verification of a given set of random numbers. The test persons are asked to alternatively verify the numbers. Both test persons get a list of randomized numbers, which are printed on a sheet of paper arranged in rows. This scenario is highly interactive and to yield high impact of transmission delay on perceived quality.

5.4.1 Test procedure

Test person A has to start reading the first row of the numbers and B has to verify the numbers in the corresponding list. Test person take turns in reading, so when A has finished reading the numbers of one row, B continues.

5.5 Results

In this section I present the results of tests performed. To do this, we will use parameters obtained subjectively (MOS) and other parameters obtained objec-

tively by some specific algorithms developed for this work. These parameters are:

- Speaker Alternation Rate (SAR)
- Double Talk (DT)
- Mutual Silence (MS)

Also, I performed an analysis of these objective parameters for each one of the different types of scenarios used, so that I can study the variation of these parameters with the different delay values in the specific context. The types of tests are:

- Random Number Verification (RNV)
- Interactive Short Conversation Test (iSCT)
- Short conversation Test (SCT).

5.5.1 General results

5.5.1.1 Mean Opinion Score (MOS)

The Mean Opinion Score (MOS) presented below were calculated as arithmetic mean value from the questionnaires the test persons answered immediately after the test.

Delay	MOS
0	4,25
200	4,32
400	4,11
800	3,81
1600	3,44

Table 5.1: MOS vs Delay table

As a first conclusion, it was noticed that there was no significant difference between the ratings of 0 ms, 200 ms and 400 ms. From delays over 800 ms we can consider that exists a threshold level that causes that the MOS drops below 4, reaching its minimum for 1600 ms. Therefore, from a subjective point of view we can conclude that the perception of the end-user is acceptable until delay levels of 800 ms or higher.

5.5.1.2 Speaker Alternation Rate (SAR)

From the results shown in the following table, we see that there is a progressive decreasement in the values of the Speaker Alternation Rate, reaching a difference of 6 points between the SAR value for delay of 200 ms and the SAR value of 1600 ms. We conclude that in the tests performed in this work, the SAR value decreases when the value of the delay increases.

Delay	SAR
0	29,29
200	31,1
400	29,42
800	25,97
1600	25,57

Table 5.2: SAR vs Delay table

5.5.1.3 Double Talk (DT)

In this case and from the results shown in the table below, we can observe that there isn't a significant variation of the Double Talk parameters when the value of delay varies.

Delay	DT
0	25,31
200	18,02
400	19,63
800	15,91
1600	17,85

Table 5.3: DT vs Delay table

5.5.1.4 Mutual Silence (MS)

The parameter of Mutual Silence (MS) has a progressive increase from a 200 ms value. As can be seen in the following table it comes to be a difference of 18 seconds between the values of MS for a delay of 200 ms and the values of MS for a delay of 1600 ms. We can conclude that in the tests conducted in this work, the value of MS increases significantly when delay is increased.

Delay	MS
0	26,10
200	24,55
400	31,12
800	40,46
1600	42,02

Table 5.4: MS vs Delay table

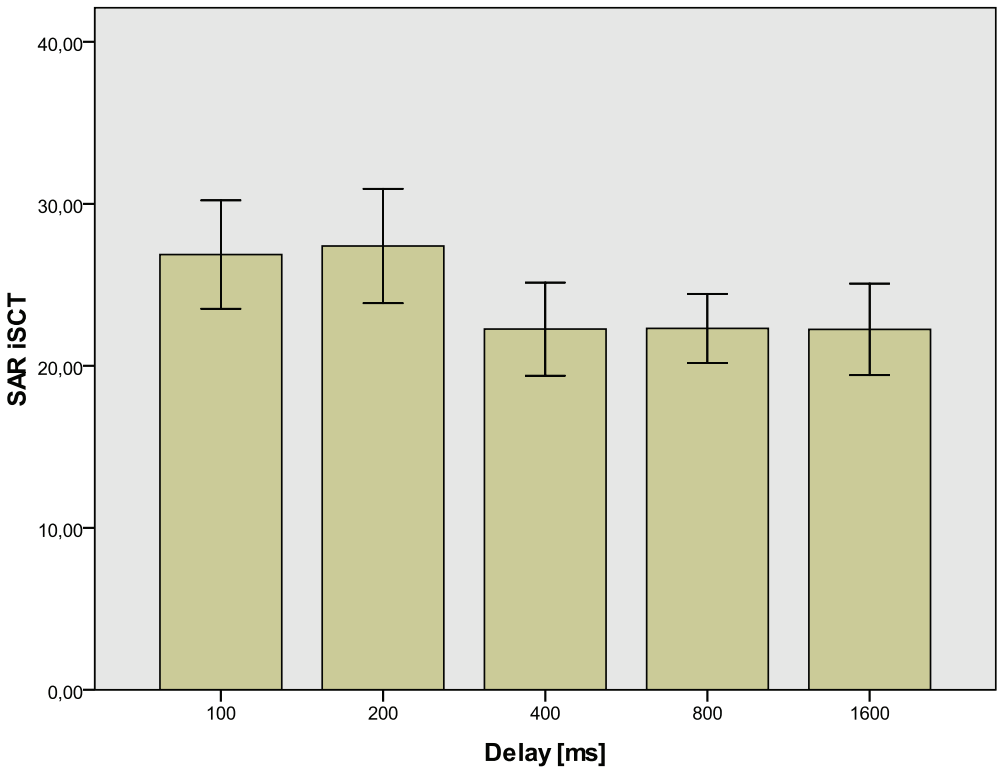
5.5.2 Scenario results

5.5.2.1 Interactive Short conversation Test (iSCT)

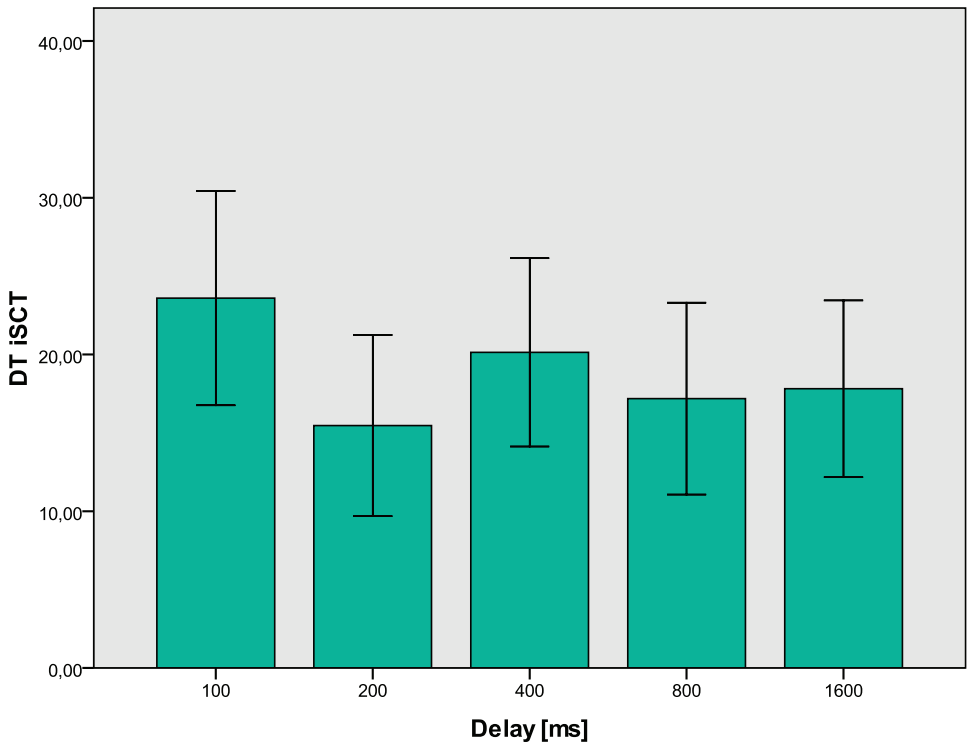
This type of scenario is characterized by more interactivity than the SCT and less than the RNV, the fact is reflected in the values of the parameters obtained. The following table shows the trend in the evolution of the objective parameters analyzed for different values of delay is similar to the case of the RNV scenarios. Specifically, for each of them:

- SAR : shows a slight decrease.
- DT : we can see a slight decrease in the value of DT.
- MS : It can be seen a significant progressive increase in the value of this parameter for increasing values of delay parameter, reaching a difference of 16 seconds between the value of MS for a delay of 0 sec and the value of MS for a delay of 1600 msec.
- MOS: shows a gradual decrease from 0 ms of delay with similar values to the general MOS (Table 5.1).
- Acceptability: has a decrease until almost seventy per cent with the highest delay.

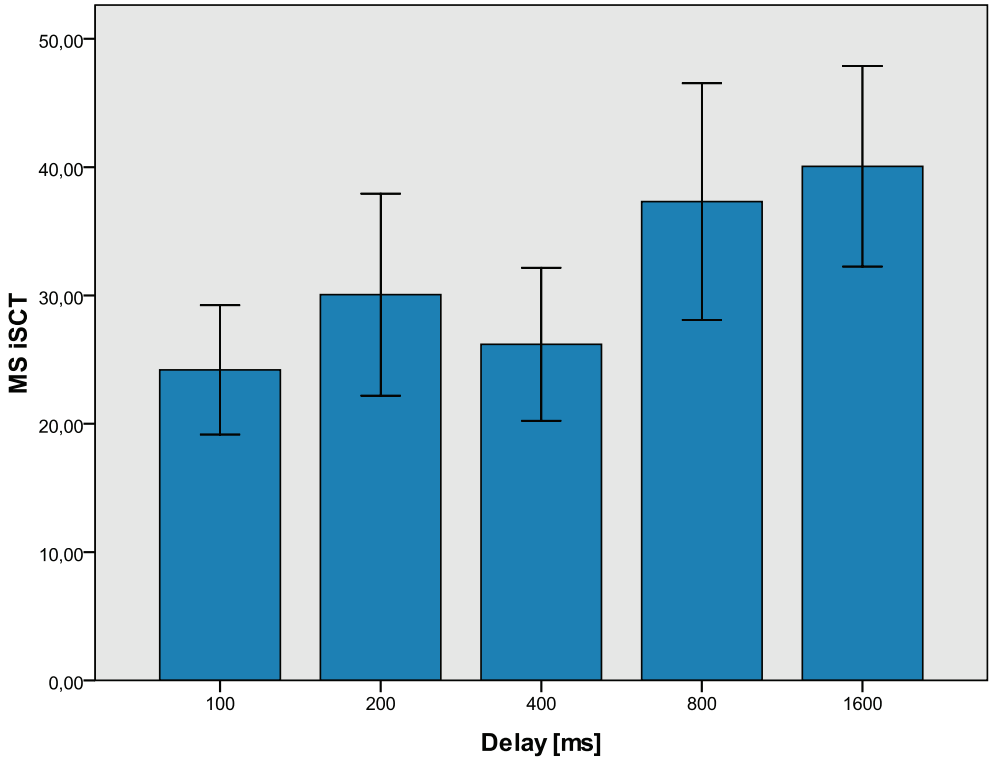
iSCT SAR results



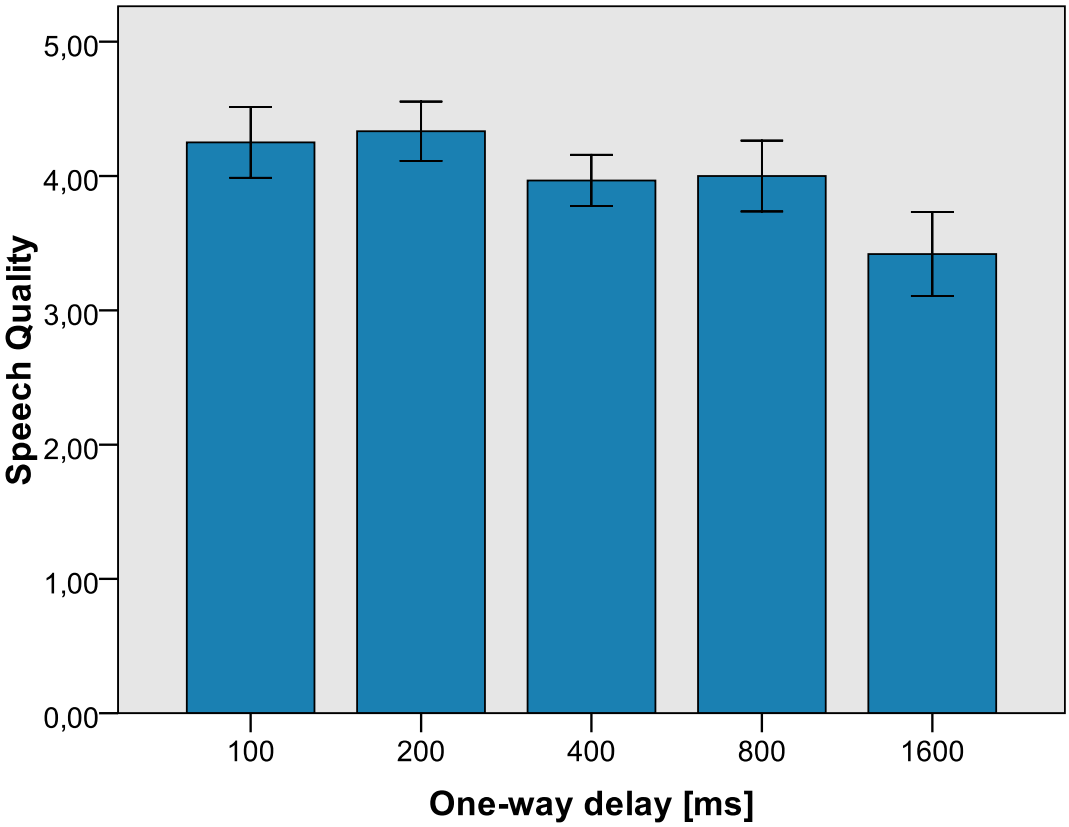
iSCT DT results



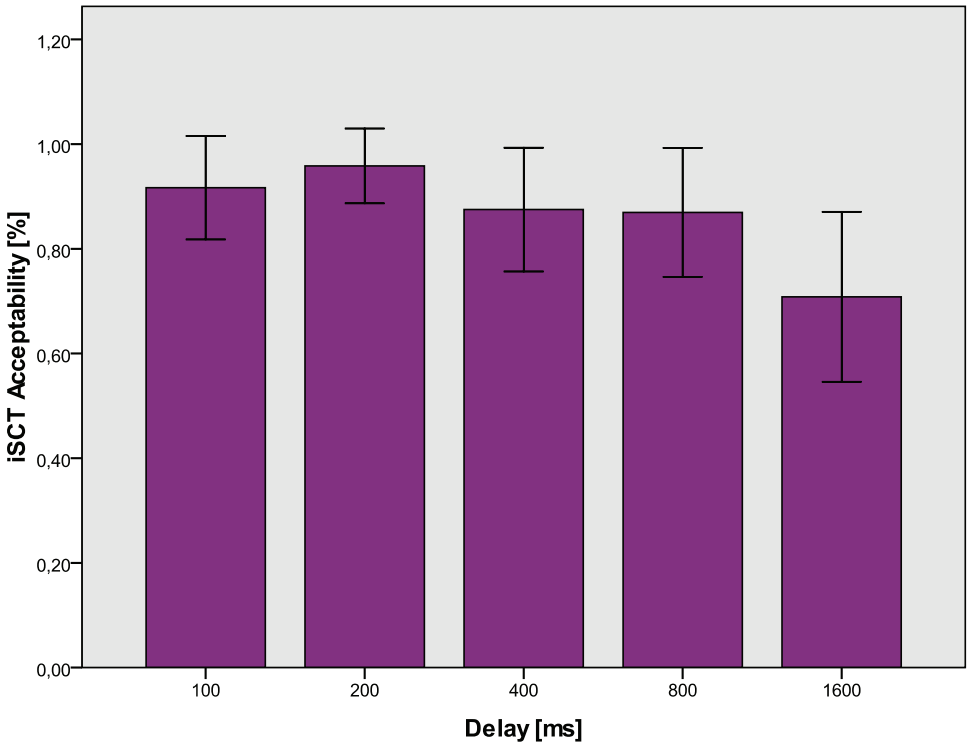
iSCT MS results



iSCT MOS results



iSCT Acceptability results



5.5.2.2 Short conversation Test (SCT)

This type of scenario is characterized by a lower interactivity than iSCT and even less than the RNV, a fact which is reflected in the values of the parameters obtained. The following table shows the evolution of the objective parameters analyzed for different values of delay. Specifically, for each of them:

SAR: no significant changes of this parameter for different values of delay.

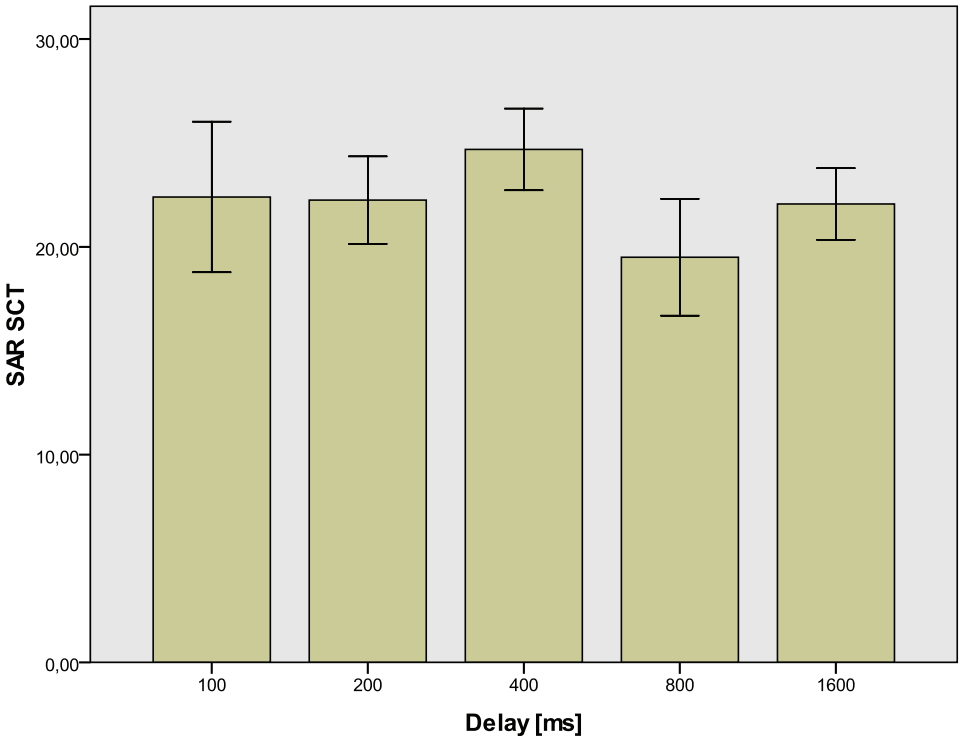
DT: no significant changes of this parameter for different values of delay.

MS: It can be seen an increase in the value of this parameter for increasing values of delay parameter, reaching a difference of 10 seconds between the value of MS for a delay of 0 sec and the value of MS for a delay of 1600 msec.

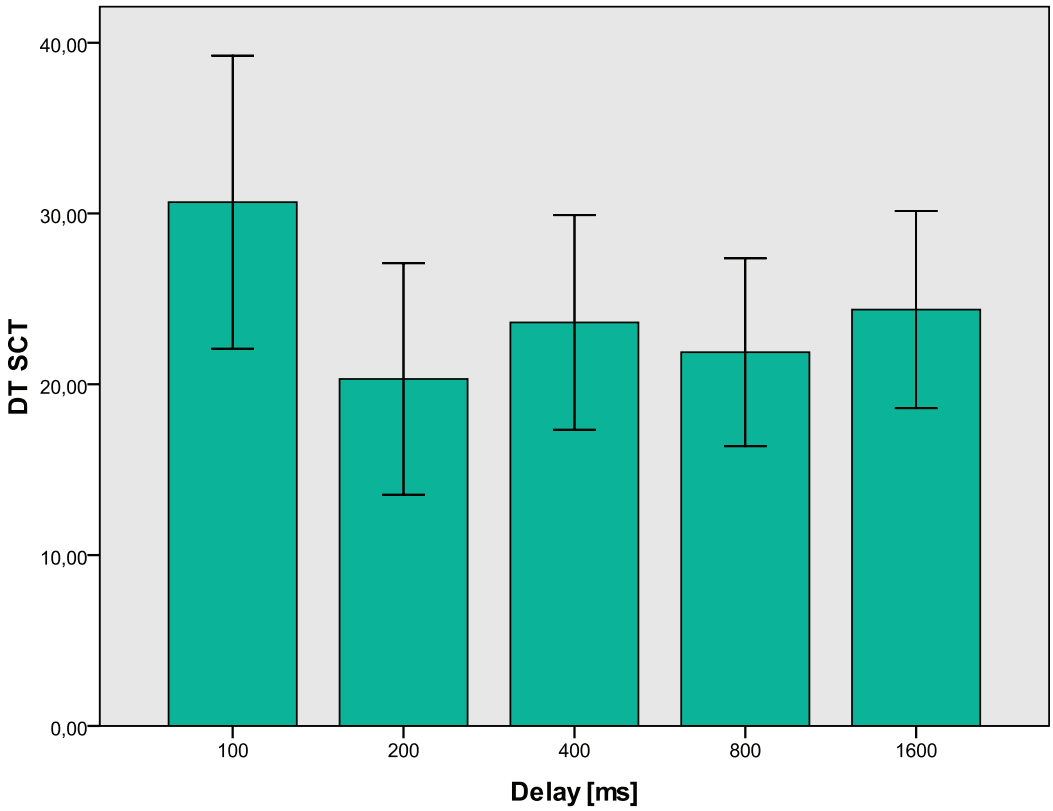
MOS: shows a slight decrease from 0 ms of delay, reaching values greater than the general MOS (Table 5.1).

Acceptability: similar decrease like in iSCT tests.

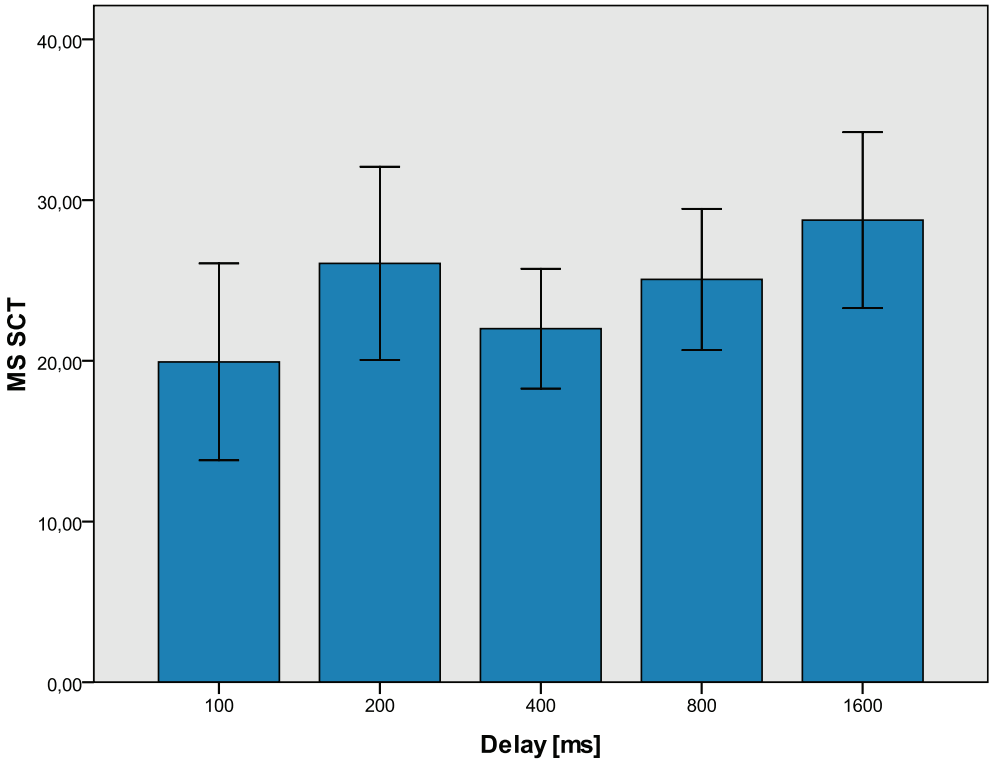
SCT SAR results



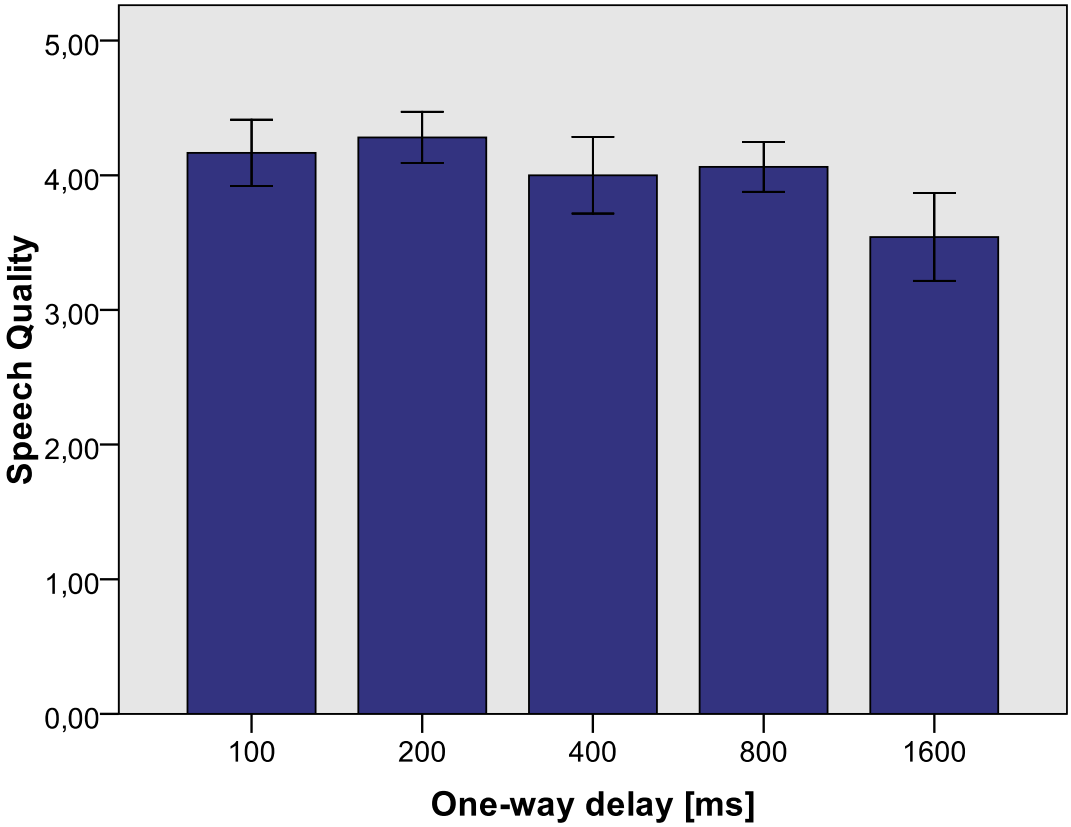
SCT DT results



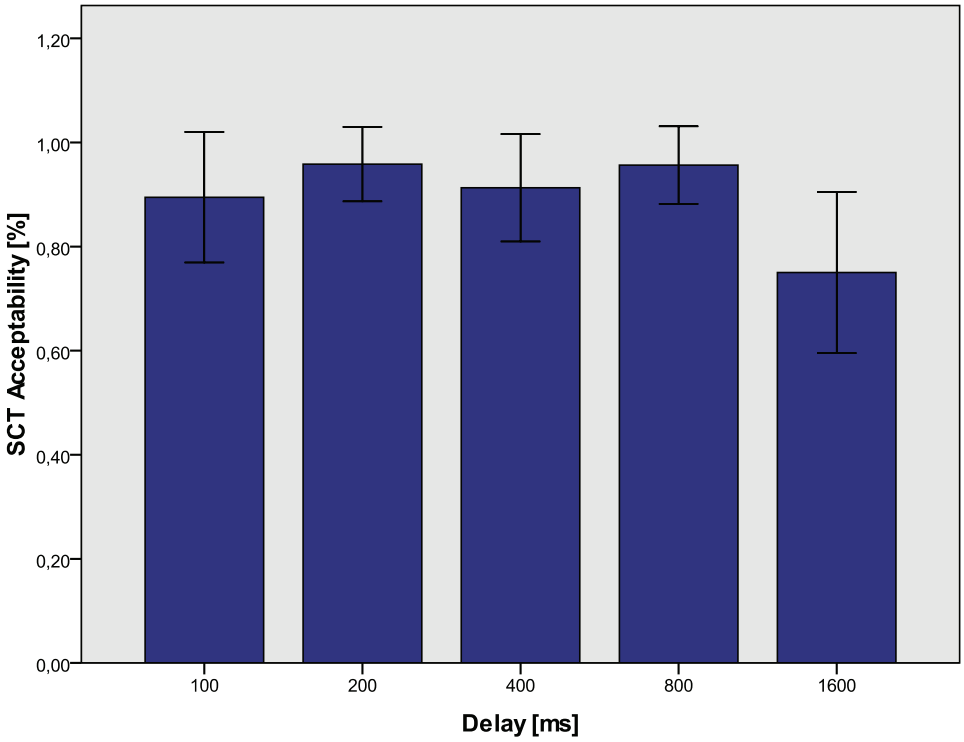
SCT MS results



SCT MOS results



SCT Acceptability results



5.5.2.3 Random Number Verification (RNV)

This type of scenario is characterized by its greater interactivity, a fact which is reflected in the values of the parameters obtained. The following table shows which has continued the trend in the evolution of the objective parameters analyzed for different values of delay. Specifically, for each one of them:

SAR: shows a gradual decrease.

DT: we can see a progressive decrease in the value of DT for increasing values of delay.

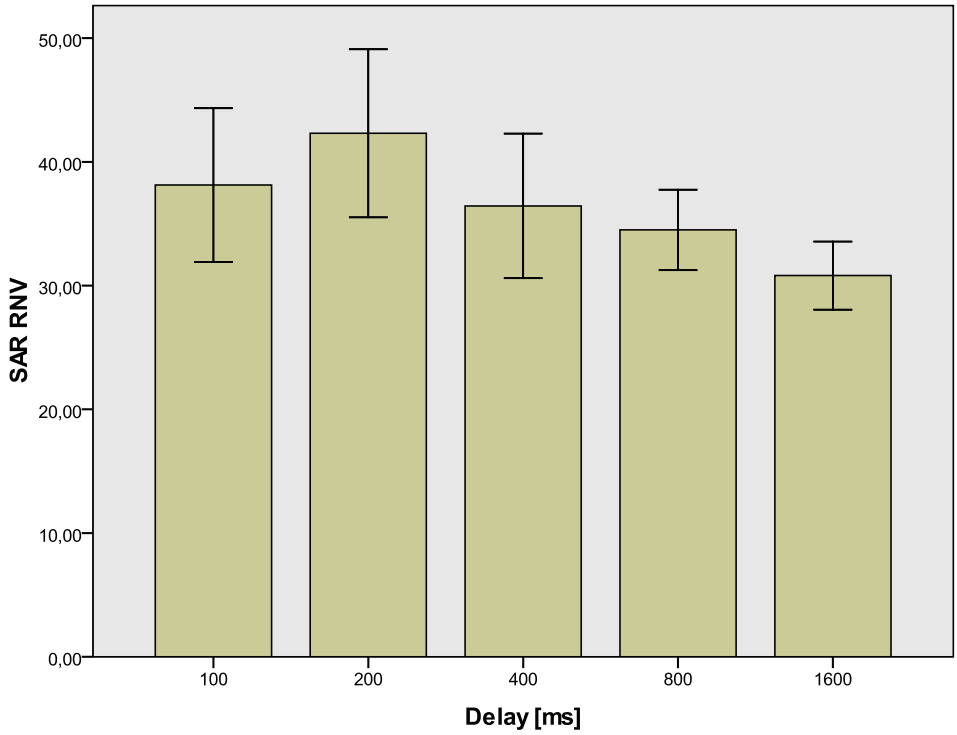
MS: it can be seen an increase in the value of this parameter for increasing values of the delay parameter, reaching a difference of 21 seconds between the value of MS for 0 seconds delay and the value for 1600 msec.

MOS: shows a gradual decrease from 0 ms of delay, reaching values minor than the general MOS (Table 5.1) from 800 ms of delay.

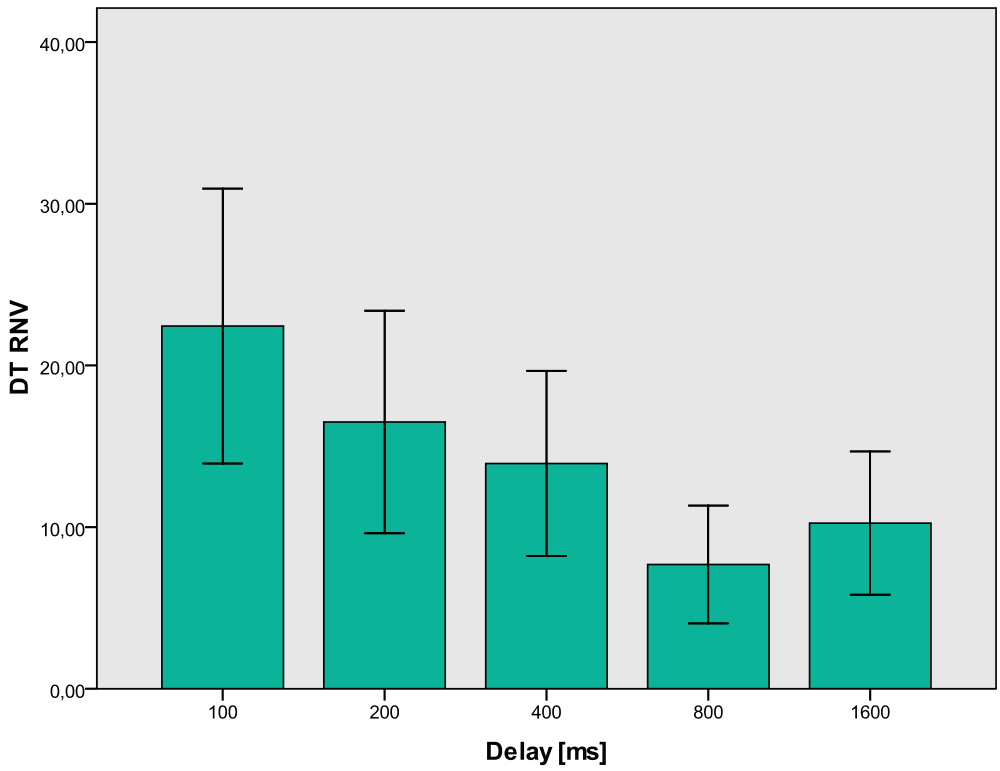
Acceptability: great decrease over 400 ms. 800 ms is in sixty per cent and 1600 ms acceptability is only forty per cent.

]

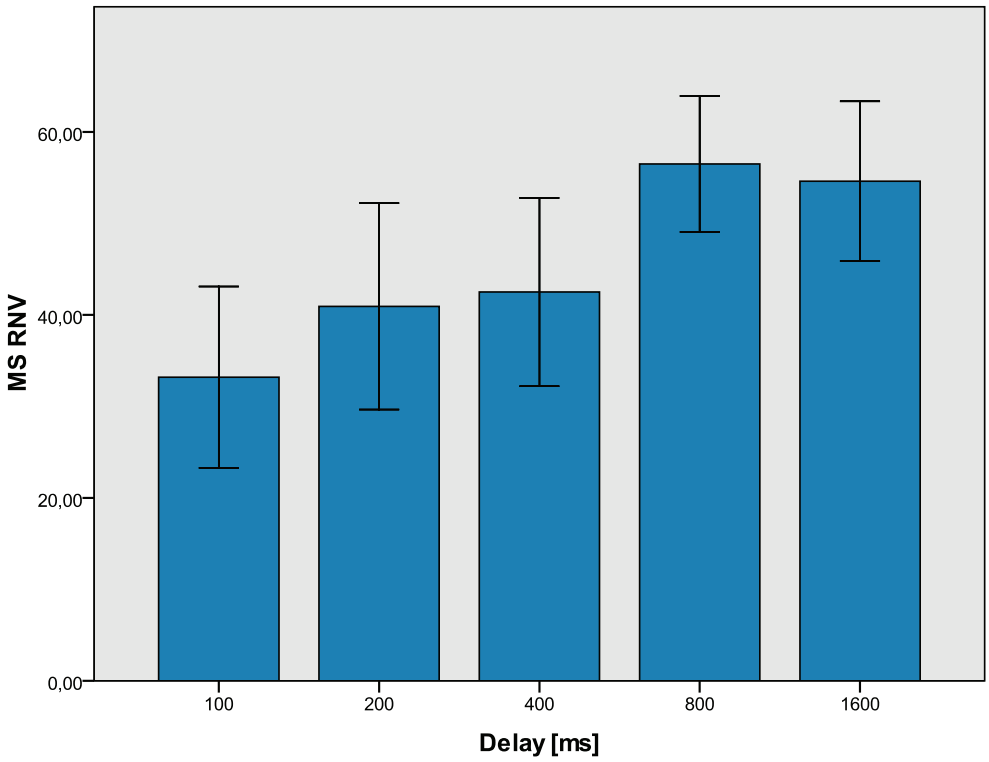
RNV SAR results



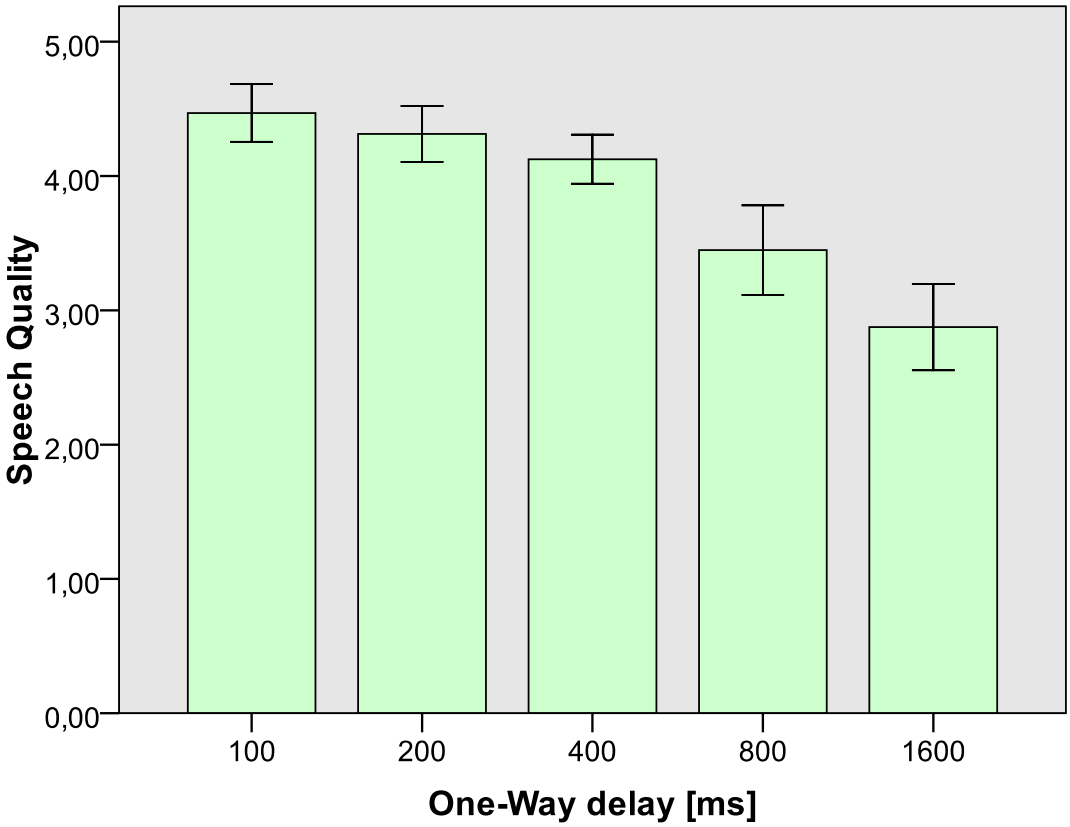
RNV DT results



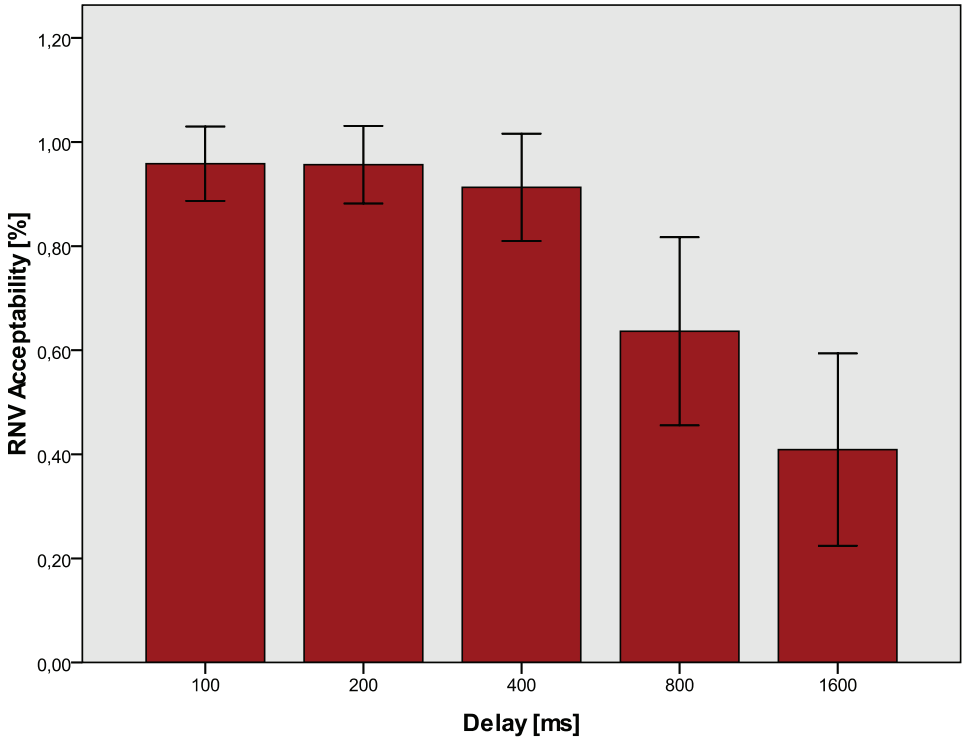
RNV MS results



RNV MOS results



RNV Acceptability results



Chapter 6

Summary and Conclusions

6.1 Summary

The number of multimedia applications running over packet based interactive real time communication system has been increasing continuously. In these communication systems, the quality is not guaranteed, especially over connections with long propagation delays. Thus, one of the major challenges is to provide to the end user acceptable levels of service quality. It's necessary to ensure that all the applications running over the network perform well, regardless of whether they are voice, video, data, real time or non real time applications.
The customer needs and expectations have to be central to product design and business process, in order to deliver an overall acceptable quality of experience (QoE) that's include security, reliability and availability.

Chapter 1 gives an introduction to features and limitations of the packet-switched networks related with the difficulties in order to guarantee the QoS, especially for the real time services like Voice over IP (VoIP). Introduces the subject quality assessment methods like a methodology that permits to know the perceived quality from the user's perspective. The second part deals with Quality of Experience (QoE) perceived by the end user when we use IP networks to transmit voice over IP (VoIP). It explains the delay impacts of these networks on the quality speech and details the individual components of the complete delay. Finally, it describes the areas of limits for one way transmission delay, included in the [7, ITU-T G.114 Recommendation].

In the first part of the Chapter 2 gives an overview about the procedure and pro-

tocols used to transmit voice over IP. It explains the major effects that impair voice quality on a VoIP systems and an introduction about the different objective and subjective measurement exist to judge the performance and QoE of VoIP service. The second part deals with the QoE and his relationship with the QoS, introducing some engineering methodology to help drive QoE into the network planning and engineering processes. The third part presents the concept of interactivity and several standardized conversation models. In the fourth part, it describes the subjective quality measurements methods that have been standardized in [1, ITU-T Rec. P.800], the test scenarios more common associated to the speech quality assessment and the different rating scales methods has been standardized. The fifth part details the interactive evaluation methods, included in [2, ITU-T Rec. P.920], for quantifying the impact of transmission delay and transmission impairments on point to point or multipoint audiovisual communications. The last part of this chapter presents different options of instrumental measurements methods that try to improve the cost and time consumed by the subjective transmission quality assessment.

Chapter 3 describes the experimental environment as it was built in FTW dependences.

Chapter 4 details the operating procedure and algorithm specifically developed by this work. In the first part explains the general procedure that has been automated. The second part included a detailed description of the Voice Activity Detection program developed. Finally, it describes the program for calculation statistical parameters.

Chapter 5 presents the general requirements for conversation test scenarios as a complement of the description included in [1, Annex A of ITU-T Rec. P.800]. The second part includes a description of the test procedure of different group of tests. At the end, it presents comments and set of values for statistical analysis.

6.2 Conclusions

The quality of packet-based telephony is influenced by several factors that may degrade the perceived speech quality. One of this is the absolute delay.

This work analyze the human to human mediated communications and how the

mediated conversation is affected by the communication defects. It deal with the delay impacts on communications and the end-user quality perception.

A test bed and conduct user tests has been built to prove assumptions from recent work stating that delay impact is less severe than the ITU-T recommendation states. The ITU-T recommends strict limits regarding the one way delay. Above 400 ms, the speech quality is supossed to be unacceptable for the users.

An experimental setup is used which allows adjustment of delay network parameter to study their impact on the perceived connection quality.

It has been developed several algorithms that permit to detect the voice activity and to calculate relevant statistics related to each scenario and network condition. These statistics are the following : Speaker Alternation Rate (SAR), Double Talk (DT) and Mutual Silence (MS). Additionaly, the test bed permit to obtain subjective parameters in order to analyze the delay effects on speech quality.

It has been designed several types of scenarios with different degrees of interactivity. This types of scenarios are the following :

1. Short Conversation Test (SCT)
2. Interactive Short Conversation Test (iSCT)
3. Random Number Verification (RNV)

In the practice test has been used five scenarios of each type and five different delays conditions. These are the following : 0 ms, 200 ms, 400 ms, 800 ms and 1600 ms. Each couple of test participants has used 15 different tests that combine delay conditions and type of test.

The practice test has been performed involving 34 test untrained persons in several sessions.

The results shows that test persons tolerate delay to a large extent. From a subjective point of view, we can conclude that the perception of the end-ser is acceptable until delay conditions of 800 ms. A delay value of 800 ms, which is double the 400 ms threshold recommended by the ITU-T for acceptable calls led to a Mean Opinion Score (MOS) of almost 4 which corresponds to a Good connection. Even

larger delays up to 1.600 ms did not result in a significant decrease of the perceived connection quality.

Also, it's possible verify that the more interactive scenarios result in higher impact of delay impairment on speech quality than less interactive scenarios. For example, for a delay condition of 1.600 ms, the difference between the MOS value in a Random Number Verification (RNV) and the MOS value in a Short Conversation Test (SCT) is more than a half point.

The results obtained of the metrics related with the conversational interactivity are the following :

- The Speaker Alternation Rate (SAR) value decreases when the value of delay increases. The difference value of SAR between delay condition of 200 ms and 1.600 ms is of 6 points.
- The Double Talk (DT) value haven't relevant variation when the value of delay varies.
- The Mutual Silence (MS) value increases when delay value is increased. The difference of MS value between delay condition of 200 ms and 1.600 ms is of 18 seconds.

The more interactive scenarios, like Random Number Verification (RNV), to emphasize the following results :

- The Speaker Alternation Rate (SAR) value decreases more quickly with the value of delay.
- The Mutual silence (MS) increases more quickly for increasing values of delay parameter.

For scenarios with a lower interactivity, like Short Conversation Test (SCT), the results are the following :

- The Speaker Alternation Rate (SAR) don't have relevant changes for different values of delay.
- The Mutual Silence (MS) increases less quickly for increases value of delay condition. The difference of MS value between delay condition of 0 ms and 1.600 ms is of 10 seconds.

In general terms, the results obtained in this work are similar to the results achieved by [5, Hammer, 2006] and [15, Brauer, 2008].

Appendix A

Algorithms

Main project program Main project

```
function vector_VAD = oscarproject_VAD(x)
%Author:        Oscar Golderos Blanco
%Email:         ogolderos@hotmail.com
%University:    Madrid Polytechnic University
%Date:          02/01/10
%Syntax:        vector_VAD = oscar_VAD(samplex);
%This function accepts an audio sample 'samplex' as input
%and returns a vector where indicates, for each segment,
%"1" if there is voice and "0" if there isn't voice.
%Also known as voice activity detection, it utilises
%the algorithm due to Rabiner & Sambur (1975)

Ini = 1;                %Initial silence duration in seconds
Ts = 0.01;              %Frame width in seconds
Tsh = 0.005;            %Frame shift in seconds
Fs = 16000;             %Sampling Frequency

ctime1 = 0;             %Time counters.
ctime2 = 0;
ctime3 = 0;

w_sam = fix(Ts*Fs);             %No of Samples/window
o_sam = fix(Tsh*Fs);            %No of samples/overlap
lengthX = length(x);            %Length of file x
segs = fix((lengthX-w_sam)/o_sam)+1;
%Number of segments in speech signal.
```

```
%Each segment is equivalent to 0.005 seconds.

sil = fix((Ini-Ts)/Tsh)+1;    %Number of segments in silent period
win = hamming(w_sam);

Limit = o_sam*(segs-1)+1;     %Start index of last segment

FrmIndex = 1:o_sam:Limit;
%Vector containing starting index for each segment.

indexi = zeros(1,lengthX);
%Three single-row vectors are created in these lines to facilitate
%computation below
indexj = indexi;

indexk = indexi;


vector_VAD = zeros(1,segs);
%Vector to hold the voice activity detection for all segments.
%It begins to all silence ("0")



%Below code computes and returns frame energy for
%all segments in speech sample
Erg_Vector = zeros(1,segs);
for u=1:segs
    nextIndex = (u-1)*o_sam+1;
    Energy = x(nextIndex:nextIndex+w_sam-1).*win;
    Erg_Vector(u) = sum(abs(Energy));
end

IMN = mean(Erg_Vector(1:sil));  %Mean silence energy (noise energy)
IMX = max(Erg_Vector);          %Maximum energy for entire utterance
I1 = 0.03 * (IMX-IMN) + IMN;    %I1 & I2 are Initial thresholds
I2 = 4 * IMN;
ITL = min(I1,I2);               %Lower energy threshold
ITU = 5 * ITL;                  %Upper energy threshold
```

```
%   *****************

NsegsITLVoice = 60;          % 90 segments (0,45 seconds)
                             % upper ITL for voice
NsegsITLSilence = 50;        % 70 segments (0,35 seconds)
                             % lower ITL for silence
NsegsITUVoice = 15;          % 20 segments (0,1 seconds)
                             % upper ITU for voice

ITLs = 1;                    % ITLs : begin upper ITL for
                             % NsegsITLVoice segments
ITLf = 1;                    % ITLf : begin lower ITL for
                             % NsegsITLSilence segments
ITUs = 1;                    % ITUs : begin upper ITU for
                             % NsegsITUVoice segments

ITLUpper = 0;                % Flag indicating ITL has been
                             % superated for NsegsITLVoice
                             % segments or more
ITUUpper = 0;                % Flag indicating ITU has been
                             % superated for NsegsITUVoice
                             % segments or more
ITLLower = 0;                % Flag indicating energy is lower
                             % than ITL for NsegsITLSilence
                             % segments or more



for i=ITLf:length(Erg_Vector)   % Until the end of vector of energy

    if (ITLUpper == 0)
        % Energy hasn't superated ITL for NsegsITLVoice segments
         if (Erg_Vector(i)>ITL)
           if ((i+NsegsITLVoice)≤length(Erg_Vector))
               % Verify the limit of vector of energy
              for j = i:i+NsegsITLVoice
                 if (Erg_Vector(j)>ITL)
               % Verify if the energy > ITL for NsegsITLVoice segments
                     ctime1 = ctime1+1;
                     indexi(ctime1) = j;
                 end
              end
              if(ctime1 ≥ NsegsITLVoice)
```

```
                ITLs = indexi(1);
                ITLUpper = 1;
                % Energy has superated ITL for NsegsITLVoice segments
                ITLLower = 0;
                ITUUpper = 0;
            end
            ctime1 = 0;
            indexi = zeros (1,lengthX);
        end
    end

else
    if (ITUUpper == 0)
        % Energy hasn't superated ITU for NsegsITUVoice segments
      if ((i+NsegsITUVoice)≤length(Erg_Vector))
          % Verify the limit of vector of energy

         if (Erg_Vector(i)>ITU)

             for j=ITLs:(ITLs+NsegsITUVoice+1)
                if (Erg_Vector(j)>ITU)
                    % Verify if the energy > ITU for
                    % NsegsITUVoice segments
                    ctime2 = ctime2+1;
                    indexj(ctime2) = j;
                end
             end
             if(ctime2 ≥ NsegsITUVoice)
                 ITUs = indexj(1);
                 ITUUpper = 1;
                 % Energy has superated ITU
                 % for NsegsITUVoice segments
             else
                 ITLUpper = 0;
                 % Begin again to search ITLs
             end
             ITLLower = 0;
             ctime2 = 0;
             indexj = zeros (1,lengthX);
         else
             if (Erg_Vector(i)<ITL)
                 ITLUpper =0;
             end
```

```
156                 end
157             end
158
159         else
160
161             if (ITLLower == 0)
162                 if (i == (length(Erg_Vector)-1))
163                     % If there isn't silence until the end
164                     % of vector from the beginning of the last
165                     % sector of voice complets with "1" until
166                     % the end of vector_VAD
167                     for j=ITLs:length(Erg_Vector)
168                         vector_VAD(j) = 1;
169                     end
170                 else
171                    if (Erg_Vector(i)< ITL)
172                      if ((i+NsegsITLSilence)≤length(Erg_Vector))
173                          %Verify the limit of vector of energy
174
175                          for j=i:i+NsegsITLSilence
176                             if (Erg_Vector(j)< ITL)
177                                 % Verify if the energy < ITL for
178                                 % NsegsITLSilence segments
179                                ctime3 = ctime3+1;
180                                indexk(ctime3) = j;
181                             end
182                          end
183                          if(ctime3 ≥ NsegsITLSilence)
184                             ITLf = indexk(1);
185                             ITLLower = 1;
186                          % Energy is under ITL for NsegsITLSilence segments
187                          end
188                          ctime3 = 0;
189                          indexk = zeros (1,lengthX);
190                      end
191                    end
192                 end
193             else
194                 if ((ITLs+(ITLf-ITLs))≤length(Erg_Vector))
195                     % Verify the limit of vector of energy
196                     for j=ITLs:ITLs+(ITLf-ITLs)
197                         % Write "1" for the segments with voice
198                         vector_VAD(j)=1;
```

```
                        % from ITLs until ITLf
                    end
                else
                    for j=ITLs:length(Erg_Vector)
                        % Write "1" for the segments with voice
                        % until the end of vector_VAD
                        vector_VAD(j)=1;
                    end
                end
                ITLLower = 0;       % Initializes the variables
                ITLUpper = 0;
                ITUUpper = 0;
                ITLs = 1;
                ITLf = 1;
                ITUs = 1;
                ctime1 = 0;
                ctime2 = 0;
                ctime3 = 0;
                indexi = zeros(1,lengthX);
                indexj = indexi;
                indexk = indexi;
            end              % of ITLLower

        end                  % of ITUUpper

    end                      % of ITLLUpper


end                          % of the general for

save ('Resultado','ITL','ITU','Erg_Vector','vector_VAD');
% Save results
```

Automat program Automatizer

```
%   CALCULATION  QUALITY OF SERVICE STATISTICS
clear all;
% UX = input('\Enter number of partner:  ');
% Placed in the directory DATA \ UX (UX pair number)
MR=zeros(15,5);
%Create and initialize the results matrix of the pair UX
w=[30;200;400;600;800;30;200;400;600;800;30;200;400;600;800];
% Matrix delays  (15*1)
MR(:,1)=w;
% Replaces first column MR by w
Nconditions = 15;
% For each condition there are two files .wav (speaker A and B)
D=dir('.');
% Build D structure with all data files  of DATA/UX directory.
% We'll have 30 files .wav
fprintf('\nSTART PROCEDURE FOR AUTOMATIC PARTNER SELECTED\n');
k=3;
for i=1:Nconditions
            % Read file name (name) of Speaker A  of the
            % structure Mx1 provided by directory ---->
            % D = dir('directory_name')
            % and give it as parameter to wavread
        D(k,1);
        % Read the file name .wav of speaker A for the condition i
        z=ans.name;
        x=wavread(z);
        % Read file name (name) of Speaker B  of the structure Mx1
        % provided by directory ----> D = dir('directory_name')
        % and give it as parameter to wavread
        D(k+1,1);
        % Read the file name .wav of speaker A for the condition i
        z=ans.name;
        y=wavread(z);
        z_oscarproject_VAD(x);
        load ('Resultado', 'vector_VAD');
        VAD_A=vector_VAD;
        z_oscarproject_VAD(y);
        load ('Resultado', 'vector_VAD');
        VAD_B=vector_VAD;
        z_statistics(VAD_A,VAD_B);
```

```
        load ('Rstatistics','NMSTOTAL','NDTTOTAL','NATTOTAL');
            % Store results in the matrix
        j=((k-1)/2);
        k=k+2;
        MR(j,2)=NATTOTAL;
            % Update results in the matrix for the condition
            % in the corresponding row (from 1 to 15)
        MR(j,3)=NDTTOTAL;
        MR(j,4)=NMSTOTAL;
        % MR(j,5)=MOS  % MOS calculation
end
fprintf ('\nFINAL PROCEDURE FOR AUTOMATIC PARTNER SELECTED\n');
```

Statistics program Statistics

```
function QoS = statistics (VAD_A, VAD_B)
%Author:        Oscar Golderos Blanco
%Email:         ogolderos@hotmail.com
%University:    Madrid Polytechnic University
%Date:          15/01/10
%Syntax:        QoS = statistics (vector_VAD_A, vector_VAD_B);
%This function compares the VAD ot two speakers,
% each one in a vector and it obtain differents
%statistics related with the mutual silence (MS), double talk (DT)
% and alternance (AT).
%The function obtain, for each statistic (MS, DT, AT),
% the value for each interval and the total accumulated.

segm = 0.005;                 % Size  of segment in seconds

lengthA = length(VAD_A);      % length of vector VAD_A
lengthB = length(VAD_B);      % length of vector VAD_B
if (lengthA ≤ lengthB)       % Select the length of shorter vector
    lengthX = lengthA;
else
    lengthX = lengthB;
end

Size = (lengthX*0.005);  % Size of vector in seconds

Voice_A = 0;         % Flags indicates voice ("1") or silence ("0")
Voice_B = 0;

Int = 10;                % Size of interval in seconds
NInt = floor(Size/Int);  % Number of intervals

NsegmInt = Int/segm;     % Number of segments in each interval

NMS = 0;                 % Number of mutual silence for each interval
NDT = 0;                 % Number of double talk for each interval
NAT = 0;                 % Number of alternate for each interval

NMSTOTAL = 0;            % Number total of MS
NDTTOTAL = 0;            % Number total of DT
NATTOTAL = 0;            % Number total of AT
```

```
vectorMS = zeros(1,NInt);   % MS in each one of interval
vectorDT = zeros(1,NInt);   % DT in each one of interval
vectorAT = zeros(1,NInt);   % AT in each one of interval


NFVA = 0;                % Number of segment begins speaker A
NFVB = 0;                % Number of segment begins speaker B

for i=1:lengthX
   if ((NFVA == 0) && (NFVB == 0))
       % If it hasn't found the beginning of any speaker
        if (VAD_A(i) == 1)
            NFVA = i;           % Beginning speaker A
        end
        if (VAD_B(i) == 1)
            NFVB = i;           % Beginning speaker B
        end
   end
end
if (NFVA ≤ NFVB)
   Voice_B = 1;                 % Begin to speak A
else
   Voice_A = 1;                 % Begin to speak B
end



for i=1:NInt                    % For each one of the intervals
    k = ((i-1)*NsegmInt)+1;
    % Index for the first segment of each interval
    for j = k:(k+NsegmInt-1)
        % For all the segments of each interval
        if (VAD_A(j) == 0)
             if (VAD_B(j) == 0)
                 NMS = NMS +1;
                 % Increase the number of mutual silence
             else
                 if ((Voice_A == 1)&&(Voice_B == 0))
                     % Change speaker A to speaker B
                     NAT = NAT+1;
                     % Increase the number of alternate
                     Voice_A = 0;
```

```
                    % Set the current condition of speaker A
                    Voice_B = 1;
                    % Set the current condition of speaker B
                end

            end


        else

            if (VAD_B(j) == 1)
                NDT = NDT +1;
                % Increase the number of double talk
            else
                if ((Voice_A == 0)&&(Voice_B == 1))
                    % Change speaker B to speaker A
                    NAT = NAT+1;
                    % Increase the number of alternate
                    Voice_A = 1;
                    % Set the current condition of speaker A
                    Voice_B = 0;
                    % Set the current condition of speaker B
                end

            end

        end
    end

    vectorMS(i) = NMS*0.005;    % Number of MS in this interval
    vectorDT(i) = NDT*0.005;    % Number of DT in this interval
    vectorAT(i) = NAT;          % Number of AT in this interval
    NMS = 0;
    NDT = 0;
    NAT = 0;

end

for i=1:NInt

    NMSTOTAL = NMSTOTAL+vectorMS(i);   % Number total accumulated of MS
    NDTTOTAL = NDTTOTAL+vectorDT(i);   % Number total accumulated of DT
    NATTOTAL = NATTOTAL+vectorAT(i);   % Number total accumulated of AT
```

```
end


save ('Rstatistics','vectorMS','vectorDT','vectorAT','NMSTOTAL',
'NDTTOTAL','NATTOTAL'); % Save results
```

Appendix B

Questions

B.1 MOS scores

Table for MOS scores:

Quality	MOS-Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table B.1: MOS scores for experienced quality (cf. [1])

B.2 Questionnaires

Table for questions:

In order to have quantitatively analyzable measures regarding the users quality perception as experienced by themselves **we used electronic questionnaires to gather their subjective quality perception**. As recommended in [1, 16, ITU-T P.800 and ITU-T P.805] we used a five item MOS scale. We asked the users to utilize this scale for the following five statements:

Please rate the speech quality (1 - 5)

Please rate the listening effort (1 - 5)

Please rate the support exercise (1 - 5)

Please rate overall quality (1 - 5)

Please rate if the conversation has been aceptable for you (yes - no)

For each of these five statements the users had the possibility to chose one item from the scale depicted in Table B.1.

Appendix C

Synchronizer

C.1 Control server

```
#!/bin/sh
# the next line restarts using wish \
exec wish8.4 "$0" "$@"

package require -exact snack 2.2
# Try to load optional file format handler
catch { package require snackogg }

snack::sound s
set last 0

proc StartRec {} {

    global session conversation
      set ::estado "RECORDING"
      set GR "RECORD"
      set GUION "_"

# COMPUTER A:  localhost ... IP computer A:
    set sock [socket localhost  23454]

    puts $sock $GR$GUION$conversation$GUION$session
    close $sock

# COMPUTER B:  localhost ... IP computer B:
```

```
26      set sock [socket localhost 23454]
27
28      puts $sock $GR$GUION$conversation$GUION$session
29      close $sock
30
31  }
32  proc StopRec {} {
33
34  # COMPUTER A:  localhost ... IP computer A:
35
36      set sock [socket localhost 23454]
37      puts $sock "STOP"
38      close $sock
39
40  # COMPUTER B:  localhost ... IP computer B:
41
42      set sock [socket localhost 23454]
43      puts $sock "STOP"
44      close $sock
45
46
47      set ::state "STOP & WRITE"
48
49  }
50
51  set state "STOPPED"
52  set session "1"
53  set conversation "1"
54
55
56  pack [label .l -text ""]
57  pack [frame .f]
58  pack [button .f.a -bitmap snackRecord -command StartRec -wi 40
59  -he 20 -fg red] -side left
60  pack [button .f.b -bitmap snackStop -command StopRec -wi 40 -he 20]
61  -side left
62  pack [button .f.c -bitmap snackPlay -command {s play} -wi 40 -he 20]
63  -side left
64  pack [button .f.d -text Exit -command exit] -si left
65  pack [label .f.e -textvar state -width 30] -side left
66  pack [frame .g]
67  pack [label .g.a -text "Conversation: " -width 30]
68  pack [entry .g.b -textvar conversation -width 16] -side right
```

```
pack [frame .h]
pack [label .h.a -text "Session: " -width 30]
pack [entry .h.b -textvar session -width 16] -side right
```

C.2 Recorder A

```

#!/bin/sh
# the next line restarts using wish \
exec wish8.4 "$0" "$@"

package require -exact snack 2.2
# Try to load optional file format handler
catch { package require snackogg }

snack::debug 0

snack::sound s
socket -server Cmd 23454

proc Cmd { sock addr port } {

    global filename

    set cmd [read $sock]

      set ::state $cmd
set RET {
}

    regsub -all $RET $cmd "" pal




      if { [regexp "RECORD" $pal] } {
        s record
        set ::state $pal
        regsub -all "RECORD" $pal "" filename

    } else {
```

```
            s stop
            set froot "Test_A"
        set ext ".wav"
        set filename $froot$filename$ext

            set menssage "STOP & WRITE: "
        set ::state $message$filename
        s write $filename
    }

    close $sock

}


set estado "WAITING CONNECTION"
pack [canvas .c -width 400 -height 30]
pack [frame .f]
# pack [button .f.a -bitmap snackStop -command StopRec -width 40
-height 20] -side left
pack [button .f.b -text Exit -command exit] -side left

pack [label .f.e -textvar state -width 60] -side left

#.c create waveform 0 0 -sound s -width 400
```

C.3 Recorder B

```
#!/bin/sh
# the next line restarts using wish \
exec wish8.4 "$0" "$@"

package require -exact snack 2.2
# Try to load optional file format handler
catch { package require snackogg }

snack::debug 0

snack::sound s
socket -server Cmd 23454

```

```
14 proc Cmd { sock addr port } {
15
16     global filename
17
18     set cmd [read $sock]
19
20       set ::state $cmd
21 set RET {
22 }
23
24     regsub -all $RET $cmd "" pal
25
26
27
28
29       if { [regexp "RECORD" $pal] } {
30         s record
31         set ::state $pal
32         regsub -all "RECORD" $pal "" filename
33
34     } else {
35             s stop
36             set froot "Test_B"
37         set ext ".wav"
38         set filename $froot$filename$ext
39
40             set message "STOP & WRITE: "
41         set ::state $message$filename
42         s write $filename
43     }
44
45     close $sock
46
47 }
48
49
50 set estado "WAITING CONNECTION"
51 pack [canvas .c -width 400 -height 30]
52 pack [frame .f]
53 # pack [button .f.a -bitmap snackStop -command StopRec -width 40
54 -height 20] -side left
55 pack [button .f.b -text Exit -command exit] -side left
56
```

```
pack [label .f.e -textvar state -width 60] -side left

#.c create waveform 0 0 -sound s -width 400
```

Appendix D

Scenarios

D.1 iSCT

D.1.1 iSCT - Test 1 Pair A

D.1.2 iSCT - Test 1 Pair B

i SCT – I:LAB

Task: Date Exchange

Your name: Meteorological Institute FTW

Replace with your conversation partner the missing information. Do this for one city after another.

(The first line is an example.)

City	Meteorological data			
	Yesterday		Today	
	Temperature	Humidity	Temperature	Humidity
Wien	15,3 °C	53%	16,5°C	63%
Linz	15,2°C	78%		
Graz	16,9°C	65%		
Salzburg	20,4°C	55%		
Innsbruck	14,8°C	84%		
Bratislava	16,2°C	77%		

i SCT 1 – Library

Task: Date Exchange

Your name: Meteorological Institute Liechtenstein

Replace with your conversation partner the missing information. Do this for one city after another.

(The first line is an example.)

City	Meteorological data			
	Yesterday		Today	
	Temperature	Humidity	Temperature	Humidity
Wien	15,3 °C	53%	16,5°C	63%
Linz			18,2°C	75%
Graz			17,1°C	61%
Salzburg			22,2°C	60%
Innsbruck			15,8°C	81%
Bratislava			16,6°C	74%

D.2 SCT

D.2.1 SCT - Test 1 Pair A

D.2.2 SCT - Test 1 Pair B

SCT 1 – I:LAB

Task: Pizza Order

Order a large pizza for 2 people at Telepizza

The pizza should be vegetarian

Ingredients : _____

_____

Price : _____ **€**

Delivery to :

Pfeilgasse 3A/7

1080 Wien

☏ 347 34 20

How long does it take for the pizza to be delivered?

_____

SCT 1 – Library

Task: Pizza Order

Your Name: Telepizza

Pizzas	1 Person	2 People	4 People
Toscana (Ham, mushrooms, tomatoes, cheese)	5,- €	9,-€	17,-€
Tonno (Tuna, onions, tomatoes, cheese)	7,-€	13,-€	25,-€
Fabrizio (Salami, ham, tomatoes, cheese)	5,-€	9,-€	17,-€
Vegetaria (Spinach, mushrooms, tomatoes, cheese)	6,-€	11,-€	21,-€

Delivery to: NAME : _____

ADDRESS : _____

_____

TELEPHONE : _____

D.3 RNV

D.3.1 RNV - Test 1 Pair A

Scenario 1 / User A

You and your partner are provided a list with random numbers. Some numbers not correspond to your partner's list.
Find the wrong numbers as fast as possible by taking turns reading them line by line.
Acknowledge by saying "YES" or "NO", and cross out the wrong numbers. You will read the A rows and your partner B rows.

A	**92**	**3**	**15**	**26**	**13**	**10**	**5**
B	87	98	40	25	55	70	91
A	**59**	**1**	**83**	**82**	**61**	**19**	**39**
B	7	35	8	16	68	32	20
A	**2**	**81**	**6**	**29**	**81**	**73**	**76**
B	47	19	31	28	93	37	11
A	**77**	**91**	**50**	**23**	**11**	**88**	**27**
B	3	74	31	85	14	53	0

D.3.2 RNV - Test 1 Pair B

Scenario 1 / User B

You and your partner are provided a list with random numbers. Some numbers not correspond to your partner's list.
Find the wrong numbers as fast as possible by taking turns reading them line by line.
Acknowledge by saying "YES" or "NO", and cross out the wrong numbers. You will read the B rows and your partner A rows.

A	**92**	**5**	**15**	**26**	**13**	**10**	**1**
B	87	98	45	25	15	70	19
A	**6**	**13**	**83**	**82**	**18**	**19**	**39**
B	7	3	8	16	88	32	20
A	**2**	**81**	**63**	**29**	**81**	**34**	**76**
B	47	54	31	28	27	32	11
A	**77**	**36**	**50**	**23**	**55**	**88**	**27**
B	3	43	31	99	14	21	0

Bibliography

[1] *ITU-T Rec. P.800 "Methods for subjective determination of transmission quality". International Telecommunication Union.*, 1996.

[2] *ITU-T Rec. P.920 "Interactive test methods for audiovisual communications". International Telecommunication Union.*, 1996.

[3] *ITU-T Rec. E.800 "Terms and definitions related to quality of service and network performance including dependability". International Telecommunication Union.*, 1994.

[4] *ITU-T Study Group 12*, 17-19 June 2008.

[5] F. Hammer, "Quality aspects of packet-based interactive speech communication," Ph.D. dissertation, Signal Processing and Speech Communication Laboratory, Faculty of Electrical and Information Engineering, University of Technology Graz, Graz, Austria, June 2006.

[6] *ITU-T Rec. G.729 "Coding of speech at 8 kbit/s using conjugate structure algebraic code excited linear predictions". International Telecommunication Union.*, 1996.

[7] *ITU-T Rec. G.114 "Methods for subjective determination of transmission quality". International Telecommunication Union.*, 2003.

[8] N. Kitawaki, "Pure delay effects on speech quality in telecommunications," *IEEE Journal on Selected Areas in Communications*, vol. Vol. 9 No. 4., 1991.

[9] *ITU-T Rec. P.800.1 "Mean Opinion Score (MOS) terminology". International Telecommunication Union.*, 2003.

[10] *Quality of Experience in real-time person-person communication – User based Qos expressed in technical network QoS terms, Proceedings of the 19th International Symposium on Human Factors in Telecommunication, Berlin, Germany*, 2003.

[11] *ITU-T Rec. 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. International Telecommunication Union.*, 2001.

[12] *ITU-T Rec. P.59 "Artificial conversational speech". International Telecommunication Union.*, 1993.

[13] U. Jekosch, "Voice and speech quality perception – assessment and evaluation," *Springer, Berlin*, 2005.

[14] *ITU-T Rec. P.82 "Method for evaluation of service from the standpoint of speech transmission quality". International Telecommunication Union.*, 1988.

[15] F. Brauer, "Experimental setup for subjective evaluation of interactive multimedia quality over ip networks," Master's thesis, Technical University Graz, 2008.

[16] *ITU-T Rec. P.805 "Subjective evaluation of conversational quality"*, 2007.